Appendix A

Random Assignment and the Target Sample Sizes

This appendix describes how random assignment was conducted and the size and allocation of the sample assembled.

The Random Assignment Process

At least 15 eligible students were recruited at each grade level in the 25 after-school centers testing each intervention (math or reading), totaling to a research sample of 2,109 students in the math centers and 2,064 students in the reading centers.[1] For programmatic reasons, random assignment was conducted separately within each center, by grade level. (Statisticians call this "blocking" by center and grade.) Even though they were blocked by grade level, the random assignment process for centers took place together, in a batch.

Prior to the point of random assignment, the centers were continuously working to build their sample of students. During this process, centers were urged to identify all potential sample members, rather than a specific number of students. For this and other reasons, until the random assignment rosters were assembled and submitted for random assignment, the exact characteristics of the sample were not known. At this point in the process, the total number of applicants per grade determined the random assignment ratio needed for that center to produce the desired size of the enhanced program group.

The Allocation of the Sample Assembled

In order to assure attendance of approximately 10 students in the enhanced class on any given day, 13 students were assigned to the enhanced program group, as long as at least 21 eligible students in a grade were on the random assignment roster. If there were 15 to 20 eligible applicants in a particular grade, the first 10 random draws were assigned to the enhanced program group so that the class could have the desired minimum number. The abilities of centers to recruit eligible students differed; thus some centers within the study had grades with too few students to produce 10 enhanced program group students with a 1:1 random assignment ratio, while some had grades where there were enough students on the random assignment roster to produce 13 students for the enhanced program group and 13 for the regular program group with a 1:1 ratio. In three cases where there were fewer than 15 students in a grade, students were as-

[1]There are three exceptions to this. One center was able to recruit only 13 third-graders and 13 fifth-graders, and another center was able to recruit only 14 second-graders.

signed in a way that maintained the ratio of two enhanced program group students for every regular program group student.

In instances where the proportion of enhanced program group students to regular program group students differs from 1:1, the power of the sample to detect impacts decreases. To compensate for the smaller sites with 15 students per grade and a 2:1 ratio, larger sites were used to increase the sample size back up to an average of 80 students per center and to move back toward the desired 1:1 ratio. To reflect the random assignment design and control for variance between blocks that random assignment produced, the random assignment block indicators are included as variables in each of the analyses.

Appendix Table A.1 shows the random assignment strategy (the number of enhanced and regular program group students) used for different numbers of students in a grade. If 21 eligible students applied, 13 students were allocated to the enhanced program group, and 8 students were assigned to the regular program group. If 26 students applied, there was a balanced design of 13 enhanced program group students and 13 regular program group students. If 32 students had applied, 13 of them would have been assigned to the enhanced program group, and 19 would go to the regular program group. More than 19 students are not assigned to the regular program group, since it would push the ratio of enhanced program group to regular program group students too far away from the ideal balanced 1:1 design with little increase in statistical precision. However, no sites had more than 32 students in a grade who were available for the study.[2]

The Evaluation of Academic Instruction in After-School Programs

Appendix Table A.1

Planned Random Assignment Ratios Given Varying Numbers of Enrolled Students

Students Enrolled per Grade, per Center	Students Randomly Assigned to Enhanced Program Group	Students Randomly Assigned to Regular Program Group
13	8	5
14	9	5
15-20	10	Remainder
More than 20	13	8-19

[2]In a few cases, exceptions to these rules were made. For example, in one district there was funding for only two teachers to work with regular program group students across all grades. In order to keep the regular program group classes to a manageable size, from a pool of 18 eligible students in a given grade, 12 were allocated to the enhanced program group, and 6 were allocated to the regular program group. Regardless of exceptions, the ratio never went beyond the worst-case scenario of 2:1.

Appendix B

Statistical Precision and Minimum Detectable Effect Size

This appendix reviews the statistical power analysis of the Evaluation of Enhanced Academic Instruction in After-School Programs impact study to determine an acceptable level of precision when estimating the impact of the program. Specifically, it reviews how the sample configuration, use of regression covariates, and other analytic assumptions would affect the precision of the impact estimates. The discussion focuses on achievement test score outcomes because of their prominence in the study.

In the discussion that follows, precision is reported as "minimum detectable effect size" (MDES). Intuitively, a minimum detectable effect is the smallest program impact that could be estimated with confidence given random sampling and estimation error.[1] This metric, which is used widely for measuring the impacts of educational programs, is defined in terms of the standard deviation of student achievement for the underlying population. For example, an MDES of 0.20 indicates that an impact estimator can reliably detect a program-induced increase in student achievement that is equal to or greater than 0.20 standard deviation of the existing student distribution. This is equivalent to approximately four Normal Curve Equivalent (NCE) points on a nationally norm-referenced achievement test and translates roughly into the difference between the 25th and the 31st percentiles.

The discussion that follows presents the smallest impact that the evaluation can reliably detect in effect size. The calculations of MDES for this study account for both within-site and across-site variation in the outcome in question. They also account for random variation across the enhanced program group and the regular program group by including pre-random assignment target test scores (reading or math). Finally, the minimum detectable differences presented here are assumed to be fixed-effect estimates; that is, they do not account for variation across sites in the true impact of the program.[2] This final assumption is justified by the fact that the sites for the study were selected purposefully. Therefore, the results are not generalizable statistically to any larger universe of after-school programs other than the centers included in this particular study.

The first row of each panel in Appendix Table B.1 shows the sample sizes resulting from various configurations of student subgroups for the math program sample and the reading program sample separately. For these rows, the first column shows the actual total number of students in the analysis samples for each subject. Each of the following columns in the table

[1] A minimum detectable effect is defined as the smallest true program impact that would have an 80 percent chance of being detected (have 80 percent power) using a two-tail hypothesis test at the 0.05 level of statistical significance.

[2] The concluding page of this appendix explains how minimum detectable differences are estimated.

shows sample sizes for the subgroups that the study aimed to include. Dividing the full analysis sample into two subgroups according to grade level equally splits the sample and creates two subgroups with 50 percent of the sample size. Defining subgroups based on their prior achievement creates somewhat unequal subgroups, with their sizes ranging from 201 students (for the "proficient" group in the reading sample, which is 11 percent of the full analysis sample) to 1,055 students (for the "basic" group in the math sample, which is 54 percent of the full analysis sample).

The second row of each panel in Appendix Table B.1 shows how the MDES for average achievement scores would vary among sample sizes associated with various configurations of student subgroups.

To see whether there is an overall program impact for math and reading, the analysis will rely on the students in the full analysis sample. For these rows, the first column of numbers indicates that the smallest program impact that could be estimated with confidence (given random sampling and estimation error in the sample) would be 0.06 standard deviation for both math and reading.

In addition to answering questions regarding effects on the full analysis sample of students, the evaluation was designed to allow for the estimation of impacts for subgroups of students defined by pre-random assignment characteristics, including students' grade levels and baseline test scores. For the minimum detectable effect rows, the remaining columns present the estimated MDES for subgroups of students that would comprise 75 percent, 50 percent, 25 percent, or 10 percent of the intended sample. For example, for a subgroup with a quarter of the full analysis sample size (457 to 490 students), the impact estimator can reliably detect a program-induced increase in student achievement that is equal to or greater than 0.12 standard deviation of the existing student distribution.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table B.1

Sample Sizes and Minimum Detectable Effect Sizes for Math and Reading, by Varying Proportions of the Analysis Sample

	Analysis Sample	75% of the Sample	50% of the Sample	25% of the Sample	10% of the Sample
Math					
Sample size	1,961	1,471	981	490	196
Minimum detectable effect size	0.06	0.07	0.08	0.12	0.18
Reading					
Sample size	1,828	1,371	914	457	183
Minimum detectable effect size	0.06	0.07	0.09	0.12	0.20

NOTE: Calculations are based on the formula discussed in Appendix B.

Estimating the MDES

Minimum detectable differences are estimated as follows:

$$MDES = M_{N-J-12} * \sqrt{\frac{\sigma_y^2(1-R^2)}{P(1-P)(N)(\sigma_y^2+\tau_y^2)} + \frac{\omega^2}{J(\sigma_y^2+\tau_y^2)}},$$

where:

M_{N-j-12} = Calculated to be 2.8, assuming a two-tailed test with a statistical power level of 0.80 and a statistical significance level of 0.05 for a sample of J blocks and N students. This multiplier assumes that estimation will include covariates for each block and 12 additional covariates.

σ_y^2 = The (within-block) variance of the outcome in question (assumed to be 1 for the effect size calculations. By definition of effect size metric, this term does not affect the MDES).

R^2 = The explanatory power of the impact regression adjusted for pre-random assignment characteristics, that is, the proportion of the variance in y explained by the experiment and any pre-random assignment characteristics. Based on the collected data, it is assumed to be 0.6.

P = The proportion of students randomly assigned to the treatment group (which equals 0.55 for the math sample and 0.57 for the reading sample).

N = The number of students: equals 1,961 for the math full analysis sample and 1,828 for the reading full analysis sample.

J = The number of grade-center blocks in the study: equals 96 for the math sample and 100 for the reading sample.

τ_y^2 = The cross-block variance in the mean value of the outcome measure y. The variance components of total outcome test scores were estimated for both reading and math and, based on the estimates,

$$\frac{\sigma^2}{\tau^2+\sigma^2} = .51 \text{ for math and } \frac{\sigma^2}{\tau^2+\sigma^2} = 0.55 \text{ for reading.}$$

ω^2 = The cross-site variance in the true impact of the program. The minimum detectable effect sizes presented here are calculated as fixed-effects estimates; that is, they do not account for cross-site variation in the true impact of the program. Thus, ω^2 is assumed to be zero.

Appendix C

Response Rates for Outcome Measures and the Creation of the Analysis Sample

This appendix describes the response rates for the data sources and the creation of the analysis sample used in the math and reading impact analysis. First the math and reading total study samples produced by random assignment are presented. Then the different response rates for the data sources used in the impact analysis are shown. Finally, this appendix compares students who responded and are thus included in the analysis sample with those not in the analysis sample, to make sure that the creation of the analysis sample did not change the specific demographic composition of students created by random assignment.[1]

The Math Sample

The intake and random assignment process produced a full study sample of 2,108 students for the math centers. Appendix Table C.1 shows the baseline characteristics for the full study sample. The response rates within this sample for the data sources are reported in this first panel of Appendix Table C.2.

The first two rows in Appendix Table C.2 show the response rates for the key outcome measures used in the impact analysis — the follow-up SAT 10 total score and the regular-school-day teacher questionnaire. The columns within the table show the percentage of all students who responded to a given measure and the proportion of respondents who are in the enhanced and regular program groups. All response rates are above 90 percent. Ninety-four percent of students (enhanced program group or regular program group) have follow-up SAT 10 math total scores, and the response rates for the teacher questionnaire are between 98 percent and 99 percent. For each data source, there is no significant difference in response rates between the enhanced and regular after-school program groups.[2] The last two rows in the first panel of Appendix Table C.2 report the response rates for the other outcome measures used in analysis: the student survey (to measure the service contrast) and the follow-up state test score (used as a

[1] Attempts were made to collect follow-up data on all students initially randomly assigned into the study, regardless of whether the student was still attending the after-school program. Thus, response rates are not reflective of attrition but, rather, of the ability of data collection staff to gather data from students.

[2] A t-test of the difference between the response rates for each data source was conducted. Differences are not statistically significant at the 0.05 level.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table C.1

Baseline Characteristics of Students in the Math Full Study Sample

Characteristic	Full Sample	Enhanced Program	Regular Program	Estimated Difference	Estimated Difference Effect Size	P-Value for the Estimated Difference
Full study sample						
Enrollment						
2nd grade	513	288	225			
3rd grade	534	291	243			
4th grade	547	297	250			
5th grade	514	292	222			
Total	2,108	1,168	940			
Race/ethnicity (%)						
Hispanic		26.22	23.64	2.58	0.06	0.13
Black, non-Hispanic		46.27	46.19	0.08	0.00	0.96
White, non-Hispanic		21.94	24.99	-3.06	-0.07	0.05
Asian		1.03	1.24	-0.21	-0.02	0.65
Other		4.54	3.94	0.61	0.03	0.49
Gender (%)						
Male		46.83	46.90	-0.07	0.00	0.97
Average age (years)		8.65	8.68	-0.03	-0.02	0.18
Overage for grade[a] (%)		18.41	19.84	-1.44	-0.04	0.39
Free/reduced-price lunch (%)						
Eligible (among information providers)		80.39	79.63	0.76	0.02	0.63
No information provided		3.51	2.57	0.94	0.06	0.21
Average household size		1.92	1.91	0.01	0.01	0.84
Single-adult household (%)		33.45	33.65	-0.20	0.00	0.92
Mother's education level (%)						
Did not finish high school		17.98	18.60	-0.62	-0.02	0.72
High school diploma or GED certificate		34.16	31.31	2.85	0.06	0.16
Some postsecondary study		41.18	44.32	-3.14	-0.06	0.14
No information provided		6.68	5.77	0.91	0.04	0.38
SAT 10 math total scaled scores		568.76	568.66	0.10	0.00	0.94
Problem Solving		573.90	573.19	0.71	0.01	0.61
Procedures		562.55	563.21	-0.66	-0.01	0.70
Sample size (total = 2,108)		1,168	940			

(continued)

Appendix Table C.1 (continued)

Characteristic	Enhanced Program	Regular Program	Estimated Difference	Estimated Difference Effect Size	P-Value for the Estimated Difference
Grade subgroups					
Grades 2 and 3					
Overage for grade[a] (%)	13.99	15.63	-1.64	-0.04	0.45
Mother's education level (%)					
Did not finish high school	19.52	18.62	0.90	0.02	0.71
High school diploma or GED certificate	33.68	30.60	3.08	0.07	0.28
Completed some post-secondary	41.45	45.09	-3.64	-0.07	0.23
No information provided	5.35	5.69	-0.33	-0.01	0.81
SAT 10 math total scaled scores	538.49	537.47	1.02	0.02	0.56
Problem solving	543.81	543.38	0.43	0.01	0.82
Procedures	532.98	530.71	2.28	0.04	0.34
Sample size (total = 1,047)	579	468			
Grades 4 and 5					
Overage for grade[a] (%)	22.75	23.99	-1.24	-0.03	0.63
Mother's education level (%)					
Did not finish high school	16.47	18.58	-2.11	-0.05	0.37
High school diploma or GED certificate	34.63	32.02	2.62	0.06	0.37
Some postsecondary study	40.92	43.56	-2.64	-0.05	0.37
No information provided	7.98	5.85	2.13	0.09	0.17
SAT 10 math total scaled scores	598.51	599.32	-0.81	-0.02	0.67
Problem solving	603.43	602.44	0.99	0.02	0.63
Procedures	591.61	595.18	-3.57	-0.06	0.13
Sample size (total = 1,061)	589	472			
Prior-achievement subgroups					
Students scoring at below basic level					
Overage for grade[a] (%)	27.51	27.24	0.27	0.01	0.95
Mother's education level (%)					
Did not finish high school	22.30	25.98	-3.68	-0.10	0.36
High school diploma or GED certificate	39.03	30.51	8.52 *	0.18	0.05
Some postsecondary study	32.34	37.00	-4.66	-0.09	0.29
No information provided	6.32	6.51	-0.19	-0.01	0.93
SAT 10 math total scaled scores	541.61	540.18	1.43	0.03	0.26
Problem solving	548.12	545.88	2.24	0.05	0.19
Procedures	530.54	529.74	0.80	0.01	0.70
Sample size (total = 516)	269	247			

(continued)

Appendix Table C.1 (continued)

Characteristic	Enhanced Program	Regular Program	Estimated Difference	Estimated Difference Effect Size	P-Value for the Estimated Difference
Students scoring at basic level					
Overage for grade[a] (%)	17.87	19.56	-1.68	-0.04	0.47
Mother's education level (%)					
Did not finish high school	16.95	19.75	-2.80	-0.07	0.24
High school diploma or GED certificate	34.82	33.58	1.24	0.03	0.67
Some postsecondary study	40.99	40.44	0.55	0.01	0.85
No information provided	7.24	6.24	1.01	0.04	0.51
SAT 10 math total scaled scores	564.32	564.54	-0.21	0.00	0.79
Problem solving	569.78	569.47	0.31	0.01	0.78
Procedures	557.43	558.94	-1.51	-0.03	0.32
Sample size (total = 1,125)	649	476			
Students scoring at proficient level					
Overage for grade[a] (%)	10.09	11.25	-1.16	-0.03	0.75
Mother's education level (%)					
Did not finish high school	13.76	6.64	7.12 *	0.18	0.05
High school diploma or GED certificate	28.44	27.34	1.10	0.02	0.82
Some postsecondary study	52.29	61.71	-9.42	-0.19	0.08
No information provided	5.50	4.31	1.19	0.05	0.59
SAT 10 math total scaled scores	601.78	602.30	-0.53	-0.01	0.68
Problem solving	605.39	604.85	0.55	0.01	0.78
Procedures	602.30	604.58	-2.28	-0.04	0.40
Sample size (total = 404)	218	186			

SOURCES: MDRC calculations are from the Evaluation of Academic Instruction in After-School Programs a ppli- cation packet and baseline results on the Stanford Achievement Test Series, 10th ed (SAT 10) abbreviated battery.

NOTES: The estimated differences are regression-adjusted using ordinary least squares, controlling for indicators of random assignment strata. The values in the column labeled "Enhanced Program" are the observed mean for the members randomly assigned to the enhanced program group. The regular program group values in the next column are the regression-adjusted means using the observed distribution of the enhanced program group across random assignment strata as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by (*) when the p-value is less than or equal to 5 percent.

The estimated difference effect size for each characteristic is calculated as a proportion of the standard deviation of the regular program group.

F-tests were calculated for the full study sample and each subgroup sample in a regression model containing the following variables: indicators of random assignment strata, math total scaled score, race/ethnicity, gender, free-lunch status, overage for grade, mother's education, mobility, and family size. The F-values are not significant for any of the samples analyzed.

There are 32 enhanced program group students and 31 regular program group students who performed at the advanced level on the baseline SAT 10; they are excluded from the prior-achievement subgroup analysis.

[a]A student is defined as overage for grade at the time of random assignment if a student turned 8 before the start of the second grade, 9 before the start of the third grade, 10 before the start of the fourth grade, or 11 before the start of the fifth grade. This indicates that the student was likely to have been held back in a previous grade.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table C.2

Response Rates to Tests, Surveys, and Observations for Students and After-School Program Staff in the Math Study Sample

Data Source	Full Study Sample	Enhanced Program Group	Regular Program Group
Students[a]			
Key outcome measures			
Follow-up SAT 10[b] (%)	94.17	93.92	94.47
Regular-school-day teacher survey (%)	98.24	97.86	98.72
Additional outcome measures			
Student survey (%)	98.06	98.12	97.98
Follow-up state test score (%)	74.76	74.91	74.57
Full study sample size (total = 2,108)		1,168	940
After-school program staff			
Additional outcome measures			
After-school staff survey[c] (%)		89.57	NA
Interviews and observations[d] (%)		100.00	NA
Sample size[e] (total = 115)			

SOURCES: MDRC calculations are from follow-up results on the Stanford Achievement Test Series, 10th ed. (SAT 10) abbreviated battery, the Evaluation of Academic Instruction in After-School Programs regular-school-day teacher survey, student survey, and after-school staff survey.

NOTES:

[a]Response rates are calculated from the full study sample for all students in the study and separately for students in each program group.

[b]This calculation is based on responses to the total math scaled score.

[c]Response rates are not calculated for regular program staff because the total sample size of regular program staff is unknown.

[d]The research team observed enhanced group instruction by randomly selecting half (51) of the 102 Mathletics staff teaching at any point in time. Following this observation, they conducted structured interviews with them. The response rate is calculated by taking the number of interviews conducted and dividing it by 51. While 3 instructors of the regular program were observed and interviewed in 2 centers where there was reported to be some structured academic instruction in math, they were not randomly selected, and thus there was no attempt to calculate a response rate for them to this measure.

[e]This is the total number of staff teaching Mathletics over the course of the school year. At a given point in time, 102 staff were teaching classes.

supplementary measure of student's academic performance).[3] Neither of these measures has a statistically significant difference in response rates between the enhanced and the regular after-school program groups. The second panel in Appendix Table C.2 presents the response rates for enhanced program staff measures, such as the after-school staff survey or the interviews and observations.

To keep the sample of students consistent across key outcome measures, an analysis sample was created to contain the students with data from both the follow-up SAT 10 achievement test score and the teacher survey. The flow chart in Appendix Figure C.1 reports the sample sizes of the analysis sample used in the impact analysis. As shown, 19 students are excluded from the math analysis sample because they have a SAT 10 score but no teacher survey; 110 students are excluded because they have a teacher survey but no SAT 10 score; and 18 are excluded because they have neither source of follow-up data. The analysis sample is 93 percent of the full study sample, and the ratio of analysis sample as a proportion of the full study sample is not statistically different between the enhanced program group and the regular program group.[4]

Even though the proportion of students included in the analysis sample is respectably high by social science research standards, it is still less than 100 percent and, therefore, raises two concerns. First, does the analysis sample differ from the full study sample? Second, within the analysis sample, are the enhanced program group and the regular program group still equivalent?

The study team examined the differences in background characteristics between the analysis sample and the rest of the study sample. While the analysis sample reflects the general characteristics of the full study sample (see Appendix Table C.1 for the full study sample's background characteristics and Table 3.3 in Chapter 3 for the analysis group's baseline characteristics), an F-test comparing the students included in the analysis sample and those in the study sample but not the analysis sample indicates that there are systematic differences between them in student characteristics. For example, students are less likely to be included in the analysis sample if their families had moved in the two years prior to the start of this study. Therefore, the students in the analysis sample are not fully representative of the full study sample of 2,108 students. Some caution should be exercised when attempting to generalize the findings beyond those who are included in the impact analysis. Nevertheless, the analysis sample contains 93 percent of students in the full study sample, making the results reflective of the behavior of most of the targeted students.

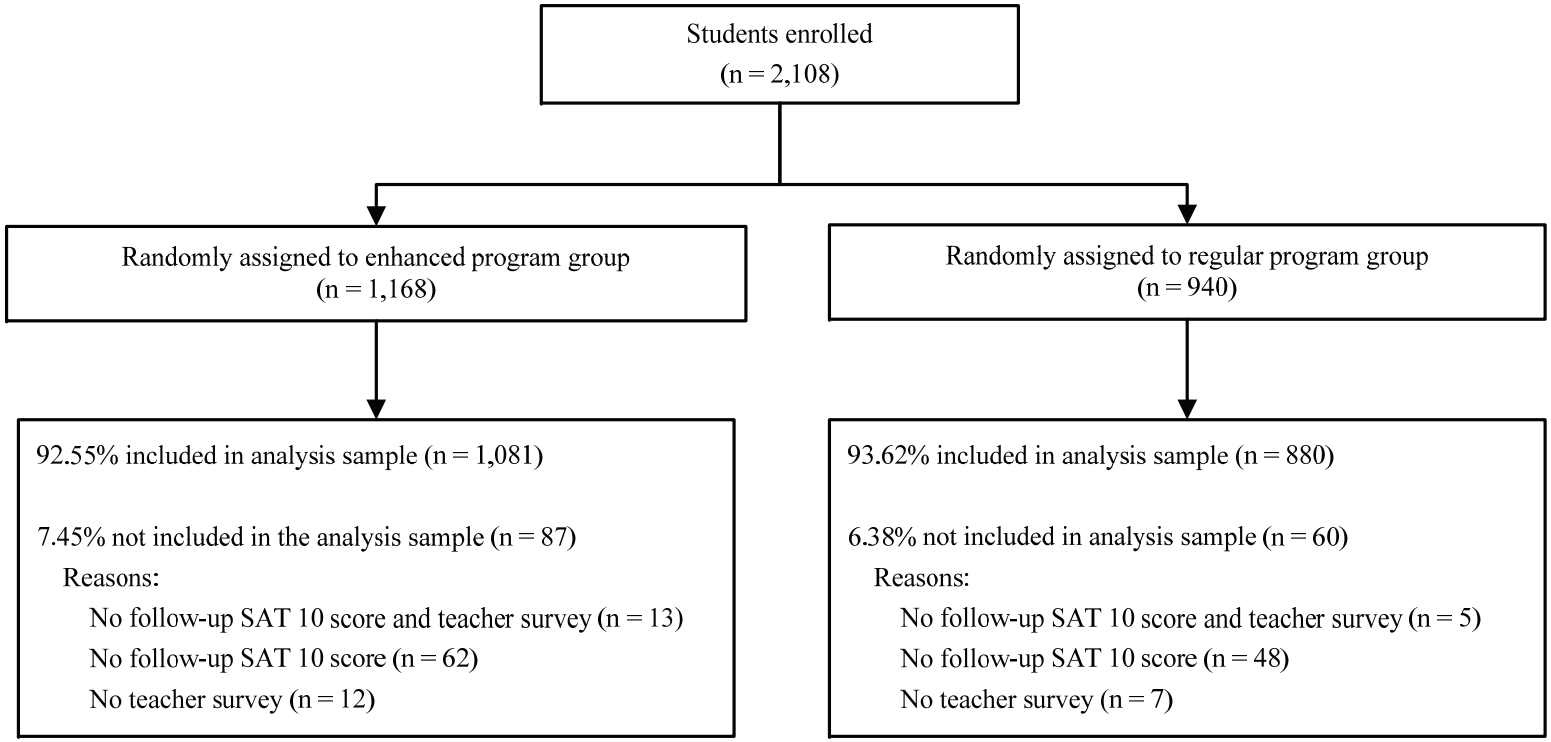
[3]Ten of the 25 schools in the math sample do not test students in grade 2, contributing to a lower response rate for this measure.

[4]Two-tailed t-tests also show that there is no significant variation in the differences in response rates between the enhanced and the regular after-school program groups across math centers, for all outcome measures and the analysis sample.

The Evaluation of Academic Instruction in After-School Programs

Appendix Figure C.1

Flow of Students from Enrollment to Math Analysis Sample



SOURCE: MDRC calculations are from the Evaluation of Academic Instruction in After-School Programs data.

NOTES: This figure explains how the math analysis sample was created from the larger group of students who enrolled in the study. All percentages are based on the number of students randomly assigned to either the enhanced or the regular program group.

In addition, Table 3.3 shows a high degree of similarity between the enhanced program group and the regular program group students in the analysis sample across the baseline characteristics. The characteristic-by-characteristic comparisons and a general F-test all indicate that, overall, there are no systematic differences between these two groups in the analysis sample. The same exercise conducted for each subgroup shows that there also are no systematic differences between the enhanced and the regular program groups at the subgroup level.

The similarity between the student characteristics of the analysis sample and the full study sample, as well as the lack of systematic differences between the enhanced and the regular program groups in the analysis sample, indicate that the analysis sample is appropriate to use in the impact analysis. This conclusion also applies to the samples of students in the subgroup analysis.

The Reading Sample

The intake and random assignment process produced a full study sample of 2,063 students for the reading centers. Appendix Table C.3 shows the baseline characteristics for the full study sample. The response rates within this sample for the data sources used in the impact analysis are reported in the first panel of Appendix Table C.4.

The first four rows in Appendix Table C.4 show the response rates for the key outcome measures used in the impact analysis: the follow-up SAT 10 reading total score, the DIBELS Oral Reading Fluency (ORF) and Nonsense Word Fluency (NWF) scores (fielded to second- and third-graders in the sample), and the regular-school-day teacher questionnaire. The columns within the table show the percentage of all students who responded to a given measure and the proportion of respondents who are in the enhanced and the regular program groups. All response rates are at or above 85 percent. As seen in the table, the response rate for both the enhanced and the regular program group students for the SAT 10 reading total score is between 91 percent and 93 percent. The response rate for both groups for the ORF test is between 88 percent and 90 percent, while the response rate for the other DIBELS portion, the NWF, is between 85 percent and 87 percent. The response rate for all groups for the teacher questionnaire can be rounded to 95 percent. For each data source, there are no significant differences in response rates between the enhanced and the regular after-school program groups.[5] The last two rows in the first panel of Appendix Table C.4 report the response rates for the other outcome measures used in the analysis: the student survey (to measure the service contrast) and the follow-up state

[5]A t-test of the difference between the response rates for each data source was conducted. Differences are not statistically significant at the 0.05 level.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table C.3

Baseline Characteristics of Students in the Reading Full Study Sample

Characteristic	Full Sample	Enhanced Program	Regular Program	Estimated Difference	Estimated Difference Effect Size	P-Value for the Estimated Difference
Full study sample						
Enrollment						
2nd grade	516	296	220			
3rd grade	524	298	226			
4th grade	526	291	235			
5th grade	497	287	210			
Total	2,063	1,172	891			
Race/ethnicity (%)						
Hispanic		23.46	24.24	-0.78	-0.02	0.57
Black, non-Hispanic		63.70	63.38	0.32	0.01	0.81
White, non-Hispanic		8.30	8.22	0.08	0.00	0.93
Asian		1.11	1.42	-0.31	-0.03	0.50
Other		3.42	2.75	0.68	0.04	0.36
Gender (%)						
Male		47.78	49.96	-2.18	-0.04	0.33
Average age (years)		8.72	8.68	0.05	0.03	0.09
Overage for grade[a] (%)		27.22	22.92	4.30 *	0.10	0.02
Free/reduced-price lunch (%)						
Eligible (among information providers)		88.13	86.24	1.89	0.06	0.16
No information provided		5.12	3.94	1.18	0.06	0.22
Average household size		1.92	1.86	0.06	0.06	0.24
Single-adult household (%)		39.53	37.66	1.88	0.04	0.38
Mother's education level (%)						
Did not finish high school		25.09	20.22	4.87 *	0.12	0.01
High school diploma or GED certificate		33.36	30.63	2.73	0.06	0.19
Some postsecondary study		37.29	43.62	-6.33 *	-0.13	0.00
No information provided		4.27	5.54	-1.27	-0.05	0.19
SAT 10 reading total scaled scores		564.36	567.66	-3.31 *	-0.08	0.01
Vocabulary/word reading[b]		554.73	559.87	-5.13 *	-0.10	0.00
Reading comprehension		565.81	569.52	-3.71 *	-0.08	0.01
Word study skills[c]		573.64	574.57	-0.93	-0.02	0.54
Sample size (total =2,063)		1,172	891			

(continued)

Appendix Table C.3 (continued)

Characteristic	Enhanced Program	Regular Program	Estimated Difference	Estimated Difference Effect Size	P-Value for the Estimated Difference
Grade subgroups					
Grades 2 and 3					
Overage for grade[a] (%)	24.58	20.31	4.26	0.10	0.09
Mother's education level (%)					
Did not finish high school	26.77	21.03	5.73 *	0.14	0.03
High school diploma or GED certificate	32.15	28.26	3.90	0.09	0.17
Some postsecondary study	37.37	44.38	-7.00 *	-0.14	0.02
No information provided	3.70	6.33	-2.63 *	-0.11	0.05
SAT 10 reading total scaled scores	536.76	541.98	-5.22 *	-0.13	0.01
Vocabulary/word reading[b]	522.02	530.41	-8.39 *	-0.16	0.00
Reading comprehension	539.02	544.46	-5.44 *	-0.12	0.01
Word study skills	551.87	554.04	-2.17	-0.05	0.31
Sample size (total = 1,040)	594	446			
Grades 4 and 5					
Overage for grade[a] (%)	29.93	25.59	4.34	0.10	0.12
Mother's education level (%)					
Did not finish high school	23.36	19.37	3.99	0.10	0.13
High school diploma or GED certificate	34.60	33.05	1.55	0.03	0.61
Some postsecondary study	37.20	42.84	-5.64	-0.11	0.06
No information provided	4.84	4.74	0.11	0.00	0.94
SAT 10 reading total scaled scores	592.53	593.90	-1.38	-0.03	0.39
Vocabulary	588.30	590.11	-1.81	-0.03	0.39
Reading comprehension	593.30	595.25	-1.95	-0.04	0.32
Word study skills[c]	595.97	595.64	0.32	0.01	0.88
Sample size (total = 1,023)	578	445			
Prior-achievement subgroups					
Students scoring at below basic level					
Overage for grade[a] (%)	33.40	31.08	2.32	0.06	0.48
Mother's education level (%)					
Did not finish high school	27.87	23.84	4.03	0.10	0.22
High school diploma or GED certificate	33.61	32.75	0.86	0.02	0.80
Some postsecondary study	33.81	37.07	-3.26	-0.07	0.34
No information provided	4.71	6.34	-1.63	-0.07	0.33
SAT 10 reading total scaled scores	546.09	547.80	-1.71	-0.04	0.09
Vocabulary/word reading[b]	531.76	535.51	-3.75 *	-0.07	0.04
Reading comprehension	546.20	548.52	-2.33	-0.05	0.11
Word study skills[c]	558.24	556.78	1.47	0.03	0.45
Sample size (total = 835)	488	347			

(continued)

Appendix Table C.3 (continued)

Characteristic	Enhanced Program	Regular Program	Estimated Difference	Estimated Difference Effect Size	P-Value for the Estimated Difference
Students scoring at basic level					
Overage for grade[a] (%)	23.57	19.45	4.12	0.10	0.13
Mother's education level (%)					
Did not finish high school	24.29	19.37	4.91	0.12	0.07
High school diploma or GED certificate	33.21	29.95	3.27	0.07	0.30
Some postsecondary study	38.39	45.56	-7.17 *	-0.14	0.03
No information provided	4.11	5.12	-1.01	-0.04	0.47
SAT 10 reading total scaled scores	573.02	574.77	-1.76 *	-0.04	0.04
Vocabulary/word reading[b]	566.01	569.44	-3.42	-0.06	0.06
Reading comprehension	574.94	577.10	-2.15	-0.05	0.13
Word study skills[c]	579.67	579.26	0.41	0.01	0.82
Sample size (total = 985)	560	425			
Students scoring at proficient level					
Overage for grade[a] (%)	20.00	6.21	13.79 *	0.33	0.01
Mother's education level (%)					
Did not finish high school	18.26	10.48	7.79	0.19	0.19
High school diploma or GED certificate	33.91	33.26	0.65	0.01	0.93
Some postsecondary study	44.35	47.35	-3.00	-0.06	0.71
No information provided	3.48	8.91	-5.43	-0.23	0.14
SAT 10 reading total scaled scores	593.99	594.81	-0.82	-0.02	0.64
Vocabulary/word reading[b]	591.46	593.93	-2.46	-0.05	0.58
Reading comprehension	598.82	600.64	-1.83	-0.04	0.58
Word study skills[c]	602.61	598.81	3.80	0.09	0.39
Sample size (total = 227)	115	112			

(continued)

SOURCES: MDRC calculations are from the Evaluation of Academic Instruction in After-School Programs application packet and baseline results on the Stanford Achievement Test Series, 10th ed (SAT 10) abbreviated battery.

NOTES: The estimated differences are regression-adjusted using ordinary least squares, controlling for indicators of random assignment strata. The values in the column labeled "Enhanced Program" are the observed mean for the members randomly assigned to the enhanced program group. The regular program group values in the next column are the regression-adjusted means using the observed distribution of the enhanced program group across random assignment strata as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by (*) when the p-value is less than or equal to 5 percent.

The estimated difference effect size for each characteristic is calculated as a proportion of the standard deviation of the regular program group.

F-tests were calculated for the full study sample and each subgroup sample in a regression model containing the following variables: indicators of random assignment strata, reading total scaled score, race/ethnicity, gender, free-lunch status, overage for grade, mother's education, mobility, and family size.

Appendix Table C.3 (continued)

The full study sample (F-value of 1.74) and the second- and third-grade sample (F-value of 1.73) are significant at the 5 percent level; the fourth- and fifth-grade sample (F-value of 1.58) is significant at the 10 percent level. The F-values for the prior-achievement subgroups are not significant.

There are 9 enhanced program group students and 7 regular program group students who performed at the advanced level on the baseline SAT 10; they are excluded from the prior-achievement subgroup analysis.

[a]A student is defined as overage for grade at the time of random assignment if a student turned 8 before the start of the second grade, 9 before the start of the third grade, 10 before the start of the fourth grade, or 11 before the start of the fifth grade. This indicates that the student was likely to have been held back in a previous grade.

[b]Second-grade students take the word reading subtest, while third- to fifth-grade students take the vocabulary subtest.

[c] The administration of the test to fifth-graders in the spring does not include word study skills.

test score (used as a supplementary measure of student's academic performance).[6] Neither of these measures has a statistically significant difference in response rates between the enhanced and the regular after-school program groups. The second panel in Appendix Table C.4 presents the response rates for enhanced program staff measures, such as the after-school staff survey or the interviews and observations.

To keep the sample of students consistent across key outcome measures, an analysis sample was created to contain the students with data from both the follow-up SAT 10 achievement test score and the teacher survey.[7] The flow chart in Appendix Figure C.2 reports the sample sizes of the analysis sample used in the impact analysis. As shown, 76 students are excluded from the reading analysis sample because they have a SAT 10 score but no teacher survey; 125 students are excluded because they have a teacher survey but no SAT 10 score; and 34 are excluded because they have neither source of follow-up data. The analysis sample is 89 percent of the full study sample, and the ratio of analysis sample as a proportion of the full study sample is not statistically different between the enhanced program group and the regular program group.[8]

[6]Thirteen of the 25 schools in the reading sample do not test students in grade 2, contributing to a lower response rate for this measure.

[7]The sample of students responding to DIBELS is unique, in that it includes only second- and third-graders. Thus, it was not used to create the reading analysis sample, nor is it limited to those students in the analysis sample. There are 96 students included in the DIBELS findings who are not part of the analysis sample: 32 of them have a SAT 10 score but no teacher survey; 53 of them have a teacher survey but no test score; and 11 have neither a SAT 10 score nor a teacher survey.

[8]Two-tailed t-tests also show that there is no significant variation in the differences in response rates between the enhanced and the regular after-school program groups across reading centers, for all outcome measures and the analysis sample.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table C.4

Response Rates to Tests, Surveys, and Observations for Students and After-School Program Staff in the Reading Study Sample

Data Source	Full Study Sample	Enhanced Program Group	Regular Program Group
Students[a]			
Key outcome measures			
Follow-up SAT 10[b] (%)	92.44	93.34	91.25
DIBELS oral reading fluency (%)	89.52	90.40	88.34
DIBELS nonsense word fluency (%)	85.96	86.70	84.98
Regular-school-day teacher survey (%)	94.67	94.71	94.61
Additional outcome measures			
Student survey (%)	96.27	96.67	95.74
Follow-up state test score (%)	74.84	75.77	73.63
Full study sample size (total = 2,063)		1,172	891
After-school program staff[c]			
Additional outcome measures			
After-school staff survey (%)		94.34	NA
Interviews and observations[d] (%)		100.00	NA
Sample size[e] (total = 106)			

SOURCES: MDRC calculations are from follow-up results on the Stanford Achievement Test Series, 10th ed. (SAT 10) abbreviated battery, results on the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) assessments, and the Evaluation of Academic Instruction in After-School Programs regular-school-day teacher survey, student survey, and after-school staff survey.

NOTES:

[a]Response rates are calculated from the full study sample for all students in the study and separately for students in each program group.

[b]This calculation is based on responses to the total reading scaled score.

[c]Response rates are not calculated for regular program staff because the total sample size of regular program staff is unknown.

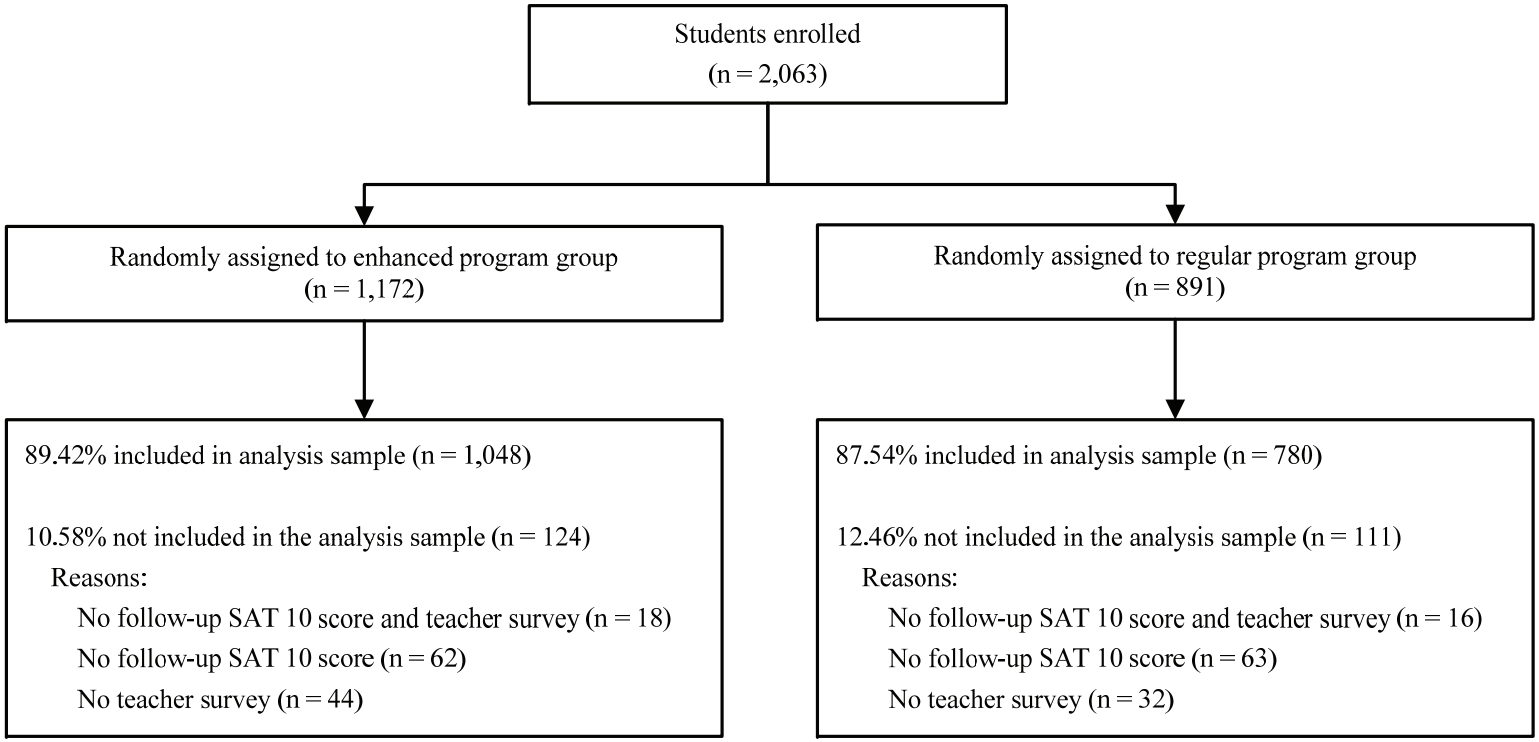
[d]The research team observed instruction by randomly selecting half (50) of the 100 Adventure Island staff teaching at any point in time; following the observation, they conducted structured interviews with them. The response rate is calculated by taking the number of interviews conducted and dividing it by 50. While 5 instructors of the regular program were observed and interviewed in 5 centers where there was reported to be some structured academic instruction in reading, they were not randomly selected, and thus there was no attempt to calculate a response rate for them to this measure.

[e]This is the total number of staff teaching Adventure Island over the course of the school year. At a given point in time, 100 staff were teaching classes.

The Evaluation of Academic Instruction in After-School Programs

Appendix Figure C.2

Flow of Students from Enrollment to Reading Analysis Sample



SOURCE: MDRC calculations are from the Evaluation of Academic Instruction in After-School Programs data.

NOTES: This figure explains how the reading analysis sample was created from the larger group of students who enrolled in the study. All percentages are based on the number of students randomly assigned to either the enhanced or the regular program group.

Similar to the math sample, even though the proportion of students included in the reading analysis sample is respectably high by standards for social science research, it is still less than 100 percent and, therefore, raises two concerns. First, does the reading analysis sample differ from the full study sample? Second, within the reading analysis sample, are the enhanced program group and the regular program group still equivalent?

The study team examined the differences in background characteristics between the analysis sample and the rest of the full study sample. While the analysis sample reflects the general characteristics of the study sample (see Appendix Table C.3 for the full study sample's background characteristics and Table 5.3 in Chapter 5 for the analysis group's baseline characteristics), an F-test comparing the students included in the analysis sample and those in the study sample but not the analysis sample indicates that there are systematic differences between them in student characteristics. For example, students are less likely to be included in the analysis sample if they are overage for grade or if information regarding family mobility prior to the start of this study is missing. Therefore, the students in the analysis sample are not fully representative of the full study sample of 2,063 students.

As discussed in Chapter 5 and shown in Table 5.3, for the reading analysis sample, differences between the enhanced program and the regular program groups on most characteristics are not statistically significant, with the exceptions being the differences in the percentage overage for grade (higher for the enhanced group), mother's education (lower for the enhanced program group), and baseline reading test scores (also lower for the enhanced program group).[9] An overall F-test across all available baseline characteristics indicates that there is a statistically significant difference at the 0.05 level between treatment and control groups for the full reading analysis. To control for these observed baseline differences, all baseline characteristics that exhibited statistically significant differences between the enhanced program and the regular program groups are included as covariates in the impact analysis model. Sensitivity tests were also conducted to ensure that the observed baseline differences do not cause selection bias in the impact analysis. (See Appendix F for details of the tests.)

As a result of these sample differences, some caution should be exercised when attempting to generalize the findings beyond students who are included in the impact analysis. Nevertheless, the analysis sample contains 89 percent of students in the full study sample, making the results reflective of the behavior of most of the targeted students.

[9]The baseline test was taken before random assignment but was scored approximately one month after the randomization. Thus, scores were not available to determine eligibility for the study or during the random assignment process.

Appendix D

Structured Protocol Observations

Observations of Implementation of Mathletics and Adventure Island

Structured protocol observations of after-school classes were conducted by local district coordinators who work on-site and were trained by Bloom Associates on the use of their respective structured protocol of implementation. These data were systematically collected to serve two purposes: (1) to provide technical assistance and (2) to describe implementation. District coordinators submitted to Bloom Associates an average of three observations for each teacher over the school year. The write-ups include a checklist of specific intended content coverage and instructional strategies of the enhanced program.

Observation forms (one for the math program and one for the reading program) were developed for this project by Bloom Associates and were reviewed by the research team and the curriculum developers, and they were used by the district coordinators during their formal observations to document whether classes used the curricular materials as intended. The protocols allow the observer to track what portions of the intended lesson are present during the class observed, what is missing entirely, and what has been modified in some way. In addition to the checklist, the write-ups on the forms document how the class was conducted, in light of the structure designed by Harcourt School Publishers or Success for All (SFA). The observation write-ups capture answers to the question "Did they do it?"

Observations of Mathletics

Appendix Box D.1 presents the guidelines for assigning points, based on which Mathletics instructional elements were recorded on the observation form as being present during the enhanced class. Bloom Associates, the curriculum developers, and the research team developed this list to summarize the observations. For the math program, a teacher could receive a maximum score of 6 points per observation by using all the instructional elements (shown in Appendix Box D.1), which include the following: sole use of the curricular materials throughout the instructional period, establishment of routines that allow for smooth transitions between the parts of the instructional session and maximizing time on task, inclusion of a teacher-led warm-up and cool-down for all students, provision of direct and differentiated instruction during the workout, use of other workout components (such as skill packs) appropriately, and inclusions of all the components in the allocated times.

Appendix Box D.1

Math Instructional Elements: Guidelines for Assigning Points

For each of the six areas listed below (uses of curriculum materials, classroom management, warm-ups and cool-downs, direct/differentiated instruction, appropriate use of other program components, structure of lesson and pacing), the district coordinator was instructed to indicate evidence of fidelity by checking bulleted items that were present. Points by area are assigned as indicated. For some of the areas, all bulleted items needed to be checked to be awarded points. In other places, an "or" indicates that only one of the bulleted items needed to be checked. Each classroom observation was recorded as a sum of the points awarded based on this protocol and point distribution scheme. NOTE: There are a total of 6 possible points for the enhanced math curriculum.

Uses curriculum materials. 1 point is awarded if:

- Observer checked box indicating students are engaged in a teacher-led Harcourt Warm-up and Cool-Down exercise;
- Observer checked box indicating the teacher provides direct instruction to small groups using page 1-2 of Skill Pack in both rotations; and
- Observer checked box indicating students work independently on the other components, such as:
    - pages 3-4 of skill packs,
    - Harcourt software connected to instruction plan, or
    - play the 24 Game and/or other Harcourt board games

[Note: A point was not given if the notes section indicated that other materials were used under any of the categories.]

Classroom management. 1 point is awarded if:

- Observer checked box indicating that during the workout portion of the class, teacher directs students to stations using established method of communication and students move quickly; or
- Notes indicate teacher uses recommended management strategies such as Popsicle sticks, rotation charts, timers, etc.

Warm-ups and cool-downs. For each, 1/2 point is awarded if:

- Observer checked box indicating students are engaged in a teacher-led or supported Harcourt numbered warm-up (or cool-down) assignment; and
- Notes indicate that all students participated (e.g., the teacher checked all students' work as she circulated…)

(continued)

Appendix Box D.1 (continued)

Direct/differentiated instruction (to individuals and small groups in rotations). 1 point is awarded if:

- Observer checked box indicating teacher provides direct instruction to small groups using pages 1 and 2 of skill pack in both rotations

Appropriate use of other components. 1 point is awarded if:

- Observer checked box indicating students moved to different activities during rotations, such as:
  - skill pack pages 3 and 4,
  - use of Harcourt software connected to the instructional plan, or
  - Harcourt board games/24 game
- When looking at the numbers of students (and their names in the notes section) assigned to component parts of the workout session, within each rotation, there is distribution across the activities mentioned above

Structure of lesson and pacing. 1 point is awarded if:

- Observer checked box indicating each component section (Warm-ups, Workout Session and Cool-downs) is completed in the allotted timeframe

Each class was observed, on average, three times during the year. For each class, observation scores were averaged together.[1] Appendix Table D.1 indicates to what extent instructional elements were present; 93 percent of classes implementing Mathletics received a score of more than 5 points, on average. In other words, a class that was observed three times may have received 5 of 6 possible points during two of the observations and 6 of 6 possible points during a third observation. The average score for that class is 5.3.

Observations of Adventure Island

Appendix Box D.2 presents the guidelines for assigning points, based on which Adventure Island instructional elements were recorded on the observation form as being present during the enhanced class. The instructional elements recorded for the reading program include slightly different components for the higher and lower reading levels, with a maximum score of

[1]Classroom scores are each teacher's mean score across all observations; when more than one teacher taught a class (for example, a teacher left the program in the middle of the year and was replaced), their mean scores are averaged together. This produces one score per grade at each center and indicates, for example, the average level of implementation that a student in a fourth-grade class at that center experienced.

Appendix Table D.1

Distribution of Structured Protocol Observation of Implementation Scores Across Mathletics Classrooms

Average Score	Percentage of Classrooms Receiving Score
Less than or equal to 1	0.00
Greater than 1 to 2	0.00
Greater than 2 to 3	0.00
Greater than 3 to 4	0.00
Greater than 4 to 5	7.45
Greater than 5 to 6	92.55
Sample size (total = 94)	

SOURCE: Structured protocol observations of implementation conducted by local district coordinators.

NOTES: Enhanced classes were observed, on average, three times during the year by district coordinators and were given a score by Bloom Associates. Classroom scores are each teacher's mean score across all observations; when more than one teacher taught a class, their mean scores are averaged together. All enhanced classes were scored on a scale of 1 to 6.

5 points per observation for Discovery Bay and Treasure Harbor classes and 6 points per observation for Alphie's Lagoon and Captain's Cove classes.[2] The instructional elements (shown in Appendix Box D.2) are a mixture of procedural factors (use of curricular materials, implementation of cooperative learning strategies, awarding of points to reward cooperative learning and the use of fluency techniques, and completion of lesson plan in the allotted time) and indicators for whether key topics were covered (phonics, fluency, and comprehension).

Each class was observed, on average, three times during the year. For each class, observation scores were averaged together.[3] Appendix Table D.2 indicates to what extent instructional elements were present. For the lower reading levels, 31 percent of classes implementing Adventure Island received a score of more than 5 points, on average. In other words, a class that

[2]Alphie's Lagoon classes (which focus on beginning-reader skills) and Captain's Cove classes (which focus on second-grade reading skills) include topics that cover phonics. Discovery Bay classes (which focus on third-grade reading skills) and Treasure Harbor classes (which focus on fourth-grade reading skills) do not include phonics as a key element.

[3]Classroom scores are calculated by taking each teacher's mean score for a specific Adventure Island level, then averaging those scores across all teachers with a score for that level at that center. This produces one score per level at each center and indicates, for example, the average level of implementation that a student in an Alphie's Lagoon class at that center experienced.

Appendix Box D.2

Reading Instructional Elements: Guidelines for Assigning Points

The Success for All (SFA) Adventure Island curriculum consists of four levels: Alphie's Lagoon, Captain's Cove, Discovery Bay, and Treasure Harbor. For each of the eight areas listed below (uses curriculum, models comprehension, completes lesson in allotted time, uses cooperative learning strategies, awards points for cooperative learning, models fluency, awards points for fluency, teaches phonics in Alphie's Lagoon and Captain's Cove), the district coordinator was instructed to indicate evidence of fidelity by checking bulleted items that were present. Points by area are assigned as indicated. For some of the areas, all bulleted items needed to be checked to be awarded points. In other places, an "or" indicates that only one of the bulleted items needed to be checked. Each classroom observation was recorded as a sum of the points awarded based on this protocol and point distribution scheme. NOTE: There are a total of 6 possible points for the Alphie's Lagoon and Captain's Cove curricula. There are a total of 5 possible points for the Discovery Bay and Treasure Harbor curricula.

Uses curriculum. 1 point is awarded if:

- Observation checklist includes name of SFA book title/day filled in on top portion; and
- Check marks assigned to relevant lesson segments and the notes sections refer to SFA curriculum as appropriate

Models comprehension. 1 point is awarded if:

- For Alphie's Lagoon, observer checked box indicating
  - story preview/review,
  - partner word and sentence reading, and
  - guided group or guided partner reading segments, when applicable
- For Captain's Cove, Discovery Bay, and Treasure Harbor, observer checked box indicating
  - the Build Background, Reading Comprehension, and Mini Lesson segments; and
  - the relevant teacher and students practice routines are highlighted or noted, such as:
    - teacher helps students make connections between their prior knowledge and the skill being taught;
    - teacher models strategy/skill;
    - teacher prompts students to review previously read text each day and make predictions, supported by evidence;
    - teacher reads aloud from the student (or secondary) text and presents additional instruction/modeling of the strategy/skill; or
    - teacher closely monitors student reading and prompts strategy use as necessary

(continued)

Appendix Box D.2 (continued)

Completes in allotted time. 1 point is awarded if:

- For all curricula,
  - the observer checks yes on the 2 prompts (1) did class begin on time and (2) timing and pacing
- For Captain's Cove, Discovery Bay, and Treasure Harbor,
  - the lesson segment check boxes (with time segments) are checked, and the notes sections do not indicate a problem with time

Uses cooperative learning strategies. 1/2 point is awarded if:

- The observer highlights or notes key words from the teacher and students practices sections of the observation protocol, such as –
  - uses Think-Pair-Share;
  - numbered heads; or
  - students actively participate in partnerships and teams

Awards points for cooperative learning. 1/2 point is awarded if:

- The observer checked box indicating "the teacher awards points for cooperation" on the Team Score Sheet section of the guide; or
- The notes section of appropriate lesson segments and/or observer comments in the general notes section at the end of the protocol indicate that cooperative learning points were awarded

Models fluency. 1/2 point is awarded if:

- In Alphie's Lagoon, the observer
  - highlights or notes key words from the teacher and student practices column of the protocol, such as —
    - teacher models fluent reading, or
    - students work with partners to read words, sentences and stories;
- In Captain's Cove, Discovery Bay, and Treasure Harbor, the observer
  - checks and/or notes key words from the sections for Partner reading and Fluency portions such as —
    - students practice fluency; or
    - teacher closely monitors practices
- In Captain's Cove, the observer checks marks in the Reading Olympics check box

Awards points for fluency. 1/2 point is awarded if:

- For all levels, the observer checks "teacher awards points for fluency"; or
- There are references in the notes sections that teacher awarded points for fluency

(continued)

Appendix Box D.2 (continued)

Teaches phonics in Alphie's Lagoon and Captain's Cove. 1 point is awarded if:

- For Alphie's Lagoon, observer checked box indicating
  - All applicable lesson segment sub-headings for the following three routines: Fast Track Phonics, Partner Word and Sentence reading, and Guided Group reading; or
  - The corresponding teacher and student practices descriptors are highlighted or referred to in notes sections
- For Captain's Cove, observer checked box indicating
  - Sail Along lesson segment; or
  - The corresponding teacher and student practices descriptors are highlighted or referred to in notes sections

was observed three times may have received 5 of 6 possible points during two of the observations and 6 of 6 possible points during a third observation. The average score for that class is 5.3. For the higher reading levels, 35 percent of classes received a score of more than 4 points, on average. In other words, a class that was observed three times may have received 4 of 5 possible points during two of the observations and 5 of 5 possible points during a third observation. The average score for that class is 4.5.

Observations of Reading and Math Instructional Practices

Observations of instructional practice were conducted by the research team using the same protocol in both math and reading sites. It is a tool developed by Public/Private Ventures (P/PV) to assess a variety of instructional variables of after-school activities. P/PV has been refining the instrument for over 10 years. P/PV has used the instrument in four previous studies of after-school programs, most recently in the CORAL (Communities Organizing Resources to Advance Learning) evaluation, which is an outcomes evaluation of an after-school literacy initiative funded by the Irvine Foundation. For the CORAL study, the instrument yielded reliable scales for such constructs as adult-youth relationships, instructional quality, and classroom management (Arbreton, Goldsmith, and Sheldon 2005).

To create the instrument, P/PV researchers reviewed both the literature on instructional practices linked to positive student learning outcomes and the after-school literature on practices linked to increased participation, to generate a set of underlying variables, or "constructs and

The Evaluation of Academic Instruction in After-School Programs

Appendix Table D.2

Distribution of Structured Protocol Observation of Implementation Scores Across Adventure Island Classrooms

Average Score	Percentage of Classrooms Receiving Score
Alphie's Lagoon and Captain's Cove classrooms	
Less than or equal to 1	2.08
Greater than 1 to 2	0.00
Greater than 2 to 3	0.00
Greater than 3 to 4	16.67
Greater than 4 to 5	50.00
Greater than 5 to 6[a]	31.25
Sample size	48
Discovery Bay and Treasure Harbor classrooms	
Less than or equal to 1	0.00
Greater than 1 to 2	7.50
Greater than 2 to 3	5.00
Greater than 3 to 4	52.50
Greater than 4 to 5	35.00
Greater than 5 to 6[a]	NA
Sample size	40

SOURCE: Structured protocol observations of implementation conducted by local district coordinators.

NOTES: Enhanced classes were observed, on average, three times during the year by district coordinators and were given a score by Bloom Associates. Classroom scores are calculated by taking each teacher's mean score for a specific Adventure Island level, then averaging those scores across all teachers with a score for that level at that center.

[a]Alphie's Lagoon classes, which focus on beginning-reader skills, and Captain's Cove classes, which focus on second-grade reading skills, are scored on a scale of 1 to 6. Discovery Bay classes, which focus on third-grade reading skills, and Treasure Harbor classes, which focus on fourth-grade reading skills, are scored on a scale of 1 to 5.

subconstructs," that seemed relevant to an after-school setting.[4] P/PV also included classroom management and adult responsiveness because those have been correlated with positive student learning outcomes (Grossman, Campbell, and Raley 2007; Miller 2006). Dimensions related to the context of the activity — such as the adequacy of the classroom space, materials, and the time allotted for completion — were also included in the observation instrument because they can affect students' ability to benefit from the activity. Finally, the observation instrument included descriptive characteristics of the activity, such as the schedule and number of adults and students present.

Constructs

The observational instrument gathers information of four overarching constructs: Instructional Delivery, Classroom Management, Cooperative Learning, and Space/Material/Time. This section describes the set of items that the team assessed to measure each construct. (The "Q" followed by a number indicates the question number for that item on the observation scales form.) The responses for all the items were done using a 4-point scale, where 1 is a low or negative rating and 4 is a high or positive rating. The following are the definitions and exact instructions that were given to observers indicating what the numbers mean:

4 = Outstanding. A score of 4 should be given when the dimension being rated is exemplary. The behaviors observed are both positive and in terms of their quality and intensity are outstanding examples of the construct; and nothing about the activity (in terms of this construct) can be improved upon. This score should be used relatively infrequently. As with all scores, ratings of 4 must be thoroughly backed up with detailed examples and descriptions of the activity along this construct.

3 = Good or very good. The activity was strong, with numerous examples of positive behaviors and no negative examples. However, while positive, the examples were not particularly outstanding. It might be helpful to think of this score as "one step down" from a score of 4 — good, but you can imagine better.

[4]Constructs are underlying variables that cannot be directly measured, such as "instruction." A construct can theoretically be made up of several subconstructs, such as organization and instructional clarity. To get a gauge — albeit an indirect gauge — of the underlying construct, a measure is created that is a collection of single-question items believed to be related to the underlying construct. (These measures are often referred to as "scales." Later, this appendix describes the scales used in this study.) This appendix uses the word "construct" to imply the underlying variables, "scale" or "measure" to indicate the indirect gauge of the construct, and "item" to specify the single question that is partially correlated with the underlying construct (DeVellis 2003).

2 = Could use improvement. There is some positive (but weak) evidence of the construct, but in contrast to a score of 3, there are also more negative examples. Significant improvement would be necessary for the activity to be considered good. A "2" may also be given in instances in which no positive behaviors are noted, if there were no negative examples either.

1 = Definitely needs improvement. There is little, if any, evidence of the construct, or predominantly negative examples. This score is also appropriate in cases where the activity is not a "bad" activity, but is simply not designed to address the construct. For example, an activity in which the adult meets with youth one-on-one without any peer interaction would receive a "1" on Peer Cooperation.

Instructional Delivery

This construct describes the manner in which the lesson is presented and its ability to create meaningful connections for youth. The construct includes the following six items.

ORGANIZATION (Q2)

This item evaluates the instructor's organization in presenting the lesson. Organization is key to successfully conveying information and instructions to youth, gaining youth's respect, and taking advantage of the limited time available in the after-school hours. Organized instructors have all materials at-hand and are prepared to start the activity on time, make efficient use of instructional time, and remain on task throughout the lesson. In assessing this item, observers must consider whether the staff appeared prepared to present the whole lesson. Did the instructors keep students focused on the activity's goals? Did they present topics with a logical sequence? On the other hand, did instructors often have to "back track" because they forgot to mention key points (making the activity seem poorly planned and disorganized)? Were they not organized enough to move smoothly from one activity to the next during the lesson?

MODELING BEHAVIOR (Q3)

This item evaluates the instructor's skill in showing students how to use the techniques being taught. In assessing this item, observers were instructed to think about whether modeling occurred during the course of the activity and whether the instructor missed obvious opportunities for modeling. When asked a question, did the teacher provide the answer or help the children think through the steps that would help them get the answer themselves?

CLARITY OF PRESENTATION (Q4)

This item assesses whether the instructor presented the goals and instructions for the activity clearly, enabling youth to move through each step of the activity without confusion. In assessing this item, observers were asked to consider the following: Did instructors explain the goals of the activity to youth in a way they could understand? Did instructors give clear and accurate directions?

Clarity of presentation is also reflected in youth's responses to the activity. Did youth know how to proceed? Did they seem confused? Were instructions provided to youth in manageable "chunks" or thrown at them in a confusing, fast-paced manner that seemed to lose them along the way? At the same time, because there may be instances in which the instructor presents materials in an extremely clear manner, yet youth are still disengaged, observers were told to base their assessment on the instructor's presentation, not the youth's response.

CONNECTION-MAKING (Q6)

This item assesses the instructor's ability to connect specific activities with other lessons and material covered and students' experiences. Successful connection-making allows youth to see the relationships between what they learn one day and the next, between their personal experiences and the material or between what they learn in school and in the after-school activity. When these connections are clear, it is easier to see why an activity is meaningful. Creating these connections also helps remind students of what they have already learned, thus making it more likely that they retain this learning.

To assess this item, observers were asked to consider the extent to which the instructor provided a context for the activity. Did s/he make connections between the current lesson and past lessons, such as explaining how the current activity relates to previous activities? Did s/he contrast or compare new information with previously learned material? Did s/he relate the current lesson to future lessons? Did s/he clearly explain how any games or activities relate to the material covered? The teacher also might ask youth what they know about the topic, referencing something in the neighborhood, or connecting the material to media or pop culture that interests the youth. Observers were instructed to assess the extent to which the instructor clearly placed each activity within the context of other material, lessons, and concepts. An activity that seemed isolated or disconnected to other material or the students' lives would score low on this item.

BALANCES INDIVIDUAL INSTRUCTION AND GROUP ACTIVITY (Q7)

There are two items in Q7. Q7a focuses on the structure of the activity and whether it was primarily an individual or group activity, as measured by the proportion of time devoted to each. Q7b focuses primarily on how well the instructor transitioned and moved between group

and individual activities. (If an activity is entirely group or individual, observers were instructed to rate Q7b as N/A since there is no transitioning between group and individual work.) Group activities were those that include the entire class. Small group and individual work were considered individual activities.

Classroom Management

This construct looks at how the instructor interacted with the students and whether the instructor managed students' behavior during the activity in ways that are appropriate for the age of youth involved and the type of activity. Successful and appropriate behavior management is essential to quality activities because it provides a positive environment for student learning (National Research Council Institute of Medicine 2004).

ADULT MANAGEMENT (Q9)

This item assesses the quality and effectiveness of the techniques staff use to manage youth behavior during the activity. How staff deal with youth who misbehave, become distracted, or disrupt the activity are key aspects of the measure. Staff's management techniques should enable the activity to precede smoothly, and at the same time, should be firm but warm. This can be displayed in a number of ways, but in all cases the adults are able to redirect the youth and win their cooperation without yelling or resorting to critical, punitive or negative discipline tactics. If behavioral issues do occur, the teacher handles them calmly and resolves them quickly and successfully. The adult handles any discipline challenges that arise without getting noticeably angry, frustrated or becoming embroiled in "power struggles" with youth. The staff may be strict with youth, but are able to correct their behavior while maintaining a positive regard and respect for the youth.

TEACHER'S INCLUSIVENESS OF YOUTH (Q10)

This item assesses the extent to which staff try to include all youth in the activity. Staff may show inclusiveness by directing questions to youth who appear isolated. Does the teacher talk to every youth at least once? Do any youth appear to be isolated, without any attention from staff?

ADULT RESPONSIVENESS (Q11)

This item assesses the quality of adult responsiveness toward students in the activity. Adult responsiveness is important for youth because it can make youth feel successful and help them benefit from the activity.

One form of adult responsiveness is the extent to which adults offer guidance to help youth understand and succeed at the task at hand, whether by providing extra information or

encouragement for youth who need it, or making themselves accessible by walking around the room or sitting at a table with youth. Adult Responsiveness includes efforts that are specifically focused on helping *all* youth to reach the goals of the activity, not just a few. Even if youth don't accept offers of help, these efforts should be noted.

MONITORING (Q5)

Teachers use monitoring techniques to assess students' progress and provide feedback. These techniques may involve asking questions to check students' understanding of the concepts being taught, circulating throughout a classroom to check on individual or group progress or providing opportunities for young people to self-assess their learning (such as checking their own work).

Within activities, a teacher must monitor both individual and group progress--and in different ways. Therefore, there are two items in Q5. Q5a focuses on how the instructor monitors and provides feedback to individuals during the direct instruction segment(s). Q5b focuses on how the instructor monitors individual progress when they are working independently and not under direct instruction (i.e., computer work, test taking, independent reading).

Cooperative Learning

Activities that are strong in peer cooperation should enable youth to interact positively with and learn from their peers. Research has shown that activities that encourage cooperative learning enhance students' desire to attend the activity more frequently (Grossman et al. 2007). This section assesses the character of the activity's peer learning environment. Q8 has two parts.

COOPERATIVE LEARNING (Q8a)

The first item focuses on the extent to which the activity requires working in pairs or collaborative problem solving (e.g., team games) and the proportion of time youth actually spend in cooperative learning activities.

MONITORING OF COOPERATIVE LEARNING (Q8b)

The second item focuses on how effectively the instructor monitors cooperative learning and actively encourages youth to work together.

Quality of Space/Material/Time

APPROPRIATENESS OF SPACE (Q12)

Dimensions of space or materials are considered under this construct — e.g., crowding, lighting, noise, quality of materials and adequacy of time. To be appropriate, an activity should not have a major problem with any dimension.

Maximizing Scoring Consistency

Observational data were collected by 16 researchers from MDRC and P/PV.

- A one-day training in Philadelphia was held for the 16 people who conducted the observations, during which time each construct and item was discussed, focusing on its behavioral indicators, how it differed from other items, and what the different rankings of scores meant.
- A scoring manual that included definitions of each item and the types of behaviors that would be positive and negative indicators was produced. The manual was distributed during the training, and observers were instructed to review it prior to conducting an observation.
- During the first two site visits, a researcher who was familiar with the instrument was paired for the same observation with a researcher who was less familiar with the instrument. Each pair rated the activity separately and then met to compare scores and resolve any discrepancies. Discussions of discrepancies served to clarify the scoring system and the definition of each item and thus increase consistency among observers.

Given the number of researchers involved in data collection, conducting traditional inter-rater reliability among all researchers was not feasible. Instead, to maximize the consistency in scoring among the group of researchers, P/PV subjected the observers' ratings of each item to review by a single P/PV researcher, who took the following steps:

- Each observer was asked to submit, along with the observation form, a detailed running record of the entire activity as well as narrative summaries of each construct.
- After the forms, narrative summaries, and running records were sent to P/PV, the P/PV researcher reviewed all the descriptive notes and ratings and compared the numerical rating scores for each item against the narrative summary of the construct and the details of the running record to check for inter-

nal rating consistency. If the numerical rating was not consistent with the narrative summary and running record, a suggested rating was written on the form by the reviewing researcher, and the form was sent back to the observing researcher with directions to review the manual again to ensure that the initial rating considered the scoring guidelines. Following this, the observing researcher provided further justification for the initial rating or accepted the change. Using a single reviewer to check each observer's ratings against the recorded details of the activity maximized consistency across observers.

Appendix E

Outcome Measures

This appendix describes the measures selected for each of the two outcome domains assessed in the study: academic achievement and academic behavior. (See Appendix Table E.1 for a summary of basic descriptive information about each outcome measure.)

Academic Achievement

At the heart of this study is a question about the impact of the enhanced after-school program on the academic achievement of students. Past evaluations, including the prior evaluation of after-school programs by Mathematica Policy Research (Dynarski et al. 2003, 2004), have relied on a nationally normed achievement test of the type used by districts or states to monitor academic performance.

Recognizing that policymakers are interested in such standardized tests, the research team, working with its Technical Work Group and the Department of Education, focused its efforts on identifying an appropriate test of math and reading for the study to administer at baseline and the end of the school year.

Study-Administered Math and Reading Test Instrument Selection

There were several criteria for selecting the achievement tests. The test used in the evaluation needed to cover grades 2 through 5 with a common framework for reporting scores and needed to have various versions, or "forms," allowing administration in both the fall (baseline) and the spring (follow-up). An effort was made to consider what tests are already being used in the study school districts and to not duplicate the testing already happening. Additionally, it was important that the test be:

1. **Accepted by the research community as a reasonable test.** In reading, there is a fairly developed view of what the key skills are for early reading (based on the National Reading Panel), and it is important that a reading test for the early grades actually measure these key skills. In math, there is not such a consensus (based on the National Mathematics Advisory Panel 2008).

2. **Seen as a policy-relevant measure of achievement.** The test should be seen as measuring the kinds of things that schools are being expected to teach and as testing them in a way that is similar to state and local accountability systems.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table E.1

Descriptive Information on Each Outcome Measure

	General Information	Norm Sample and Psychometric Properties
Stanford Achievement Test Series, 10th ed. (SAT 10) abbreviated battery	Commercially available. Math test contains two subtests: problem solving and procedures. Reading test contains three subtests: word study skills, reading comprehension, and vocabulary.	Normed to a national sample of 250,000 students in spring 2002 and of 110,000 students in fall 2002. The average student in the norm sample has a normal curve equivalent score of 50, and the standard deviation of normal curve equivalent scores is 21.06. Internal consistency (KR-20) reliability coefficients range from 0.77 to 0.95 for abbreviated multiple-choice battery test and subtests.
Dynamic Indicators of Basic Early Literacy Skills (DIBELS)	Commercially available. Contains a set of standardized, individually administered measures of early literacy development, used to monitor the development of pre-reading and early reading skills.	Benchmark and progressive goals initially were derived based on data from all schools participating in the DIBELS Data System during the 2000-2001 and 2001-2002 academic years. Test-retest reliability for elementary students ranges from 0.92 to 0.97.
State-administered tests	Norm-referenced tests are commercially available. Criterion-referenced tests are developed specifically for a state and are not commercially available. (See Appendix Tables E.2 and E.3 for a listing of the tests.)	No norming and psychometric properties are available for the criterion-referenced tests. For the norm-referenced tests: • TerraNova reading assessment: Normed to a national sample of 171,000 students. Internal consistency coefficients range from 0.76 to 0.97 for the complete battery test. • Scantron math assessment: Inter-testlet internal consistency coefficients range from 0.512 to 0.876. Correlations between individual units and overall score range from 0.747 to 0.876. Highly predictive correlation with the Iowa Test of Basic Skills as well as the Dakota State Test of Educational Progress. This computer-adaptive test stops testing the student once it reaches a reliability coefficient of 0.91.
Regular-school-day teacher survey	Questions constructed by MDRC or adapted from questions used in other after-school evaluations.[a] Survey items cover issues on homework completion and academic behavior in class.	This measurement is not nationally normed.

SOURCES: Harcourt Assessment (2004); Dynamic Indicators of Basic Early Literacy Skills (2007a); Salvia and Ysseldyke (2001); Scantron Corporation (2005); Dynarski et al. (2004).

NOTE: [a]Three single-item questions used as school-day academic behavior outcomes were drawn from the "Elementary School Teacher Survey" used for the National Evaluation of the 21st Century Community Learning Centers Program study.

3. **Feasible to administer in the after-school setting.** The realities of after-school programs and the staffing available to field the tests create some constraints in administration. Thus, the goal was to pick a test that is relatively straightforward for staff without special expertise to administer to groups of students and that takes no longer than an hour or possibly 90 minutes to administer.

4. **Scored in a way that can be combined across grades in the analysis.** In order to conduct the analysis on the full sample, the test must yield scores or measures that can be combined across grades.

5. **Sensitive to improvements at the bottom range of the achievement distribution.** The most important target group for enhanced instruction in after-school programs is students who are not doing well in school, so the goal was to pick a test that is good at picking up the changes at the low end of the distribution.

From these criteria, a list was created of possible tests, and this list was presented to the Technical Working Group, along with a memo explaining the rationale for why each test was on the list. From this list the Stanford Achievement Test, Tenth Edition (SAT 10), abbreviated battery was chosen.[1]

The SAT 10 abbreviated battery is a group-administered multiple-choice test of one hour or less. This test is widely used, nationally recognized, similar to tests that are part of state and/or local accountability systems (so it has policy relevance), and is relatively easy to administer. Based on the Technical Data Report by Harcourt:

> Stanford 10 full-length and Stanford 10 Abbreviated are both expressed on the same underlying ability scale. Although the relationship of raw score to ability may differ from one test form to another, the relationship of ability (scaled score) to percentile rank is the same. There is in essence a single norm set which applies equally to any Stanford 10 form linked to the underlying Stanford 10 scale. Thus, any information that pertains to norms for the Stanford 10 full-length test applies equally to Stanford 10 Abbreviated. Because the abbreviated form is a core subset of items on the full-length form, all of the validity information for the full-length form applies equally

[1]The SAT 10 is published by Harcourt Assessment, a sister organization of Harcourt School Publishers, which is the creator of the new math curriculum. However, the SAT 10 operates separately, and the Harcourt math curriculum is not especially aligned with the "Stanford" test.

> to the abbreviated form. The only real difference is that since the abbreviated form has fewer items, it does not measure with quite the same precision as the full-length test due to the slightly lower reliability. (Harcourt Assessment 2004, p. 46)

The SAT 10 abbreviated battery is normed to a national sample of 250,000 students in spring 2002 and of 110,000 students in fall 2003. The average student in the norm sample has a Normal Curve Equivalent (NCE) score of 50, and the standard deviation of NCE scores is 21.06. The internal consistency (KR-20) coefficients range from 0.77 to 0.95 for the abbreviated multiple-choice battery test and subtests. There is well-documented evidence of its content, criterion-related, and construct validity (Harcourt Assessment 2004). The test was administered at both baseline and follow-up, covering the topic (reading or math) addressed in the curriculum to be tested in the site.

The reliability coefficients of the abbreviated measure for the total reading score for grades 2 through 5 range from 0.90 to 0.93 for the spring test and from 0.93 to 0.95 for the fall test. For total math score, the reliability measures for grades 2 through 5 range from 0.89 to 0.92 for the spring test and from 0.88 to 0.92 for the fall test. For more details, see Appendix C of the Stanford Achievement Test Series, Tenth Edition, Technical Data Report (Harcourt Assessment 2004).

The math test contains two subtests — problem-solving and procedures — that measure content and process. Problem-solving measures the skills and knowledge necessary to solve problems in mathematics through geometry and measurement; patterns, relationships, and algebra; and data, relationships. and probability. Procedures measure the ability to apply the rules and methods of arithmetic to problems that require arithmetic solutions through computation with whole numbers, decimals, and fractions (Harcourt Assessment 2007).

The reading test contains three subtests — word study skills, reading comprehension, and vocabulary — that reflect and support a balanced, developmental curriculum and sound instructional practices. Word study skills measures structural and phonetic analysis, such as identifying and decoding compound words and contractions and recognizing sounds of consonants and vowels. Reading vocabulary measures students' understanding of the printed word, synonyms, and multiple-meaning words. Reading comprehension measures students' initial understanding, interpretation, and critical analysis of reading passages (Harcourt Assessment 2007).

Study-Administered Fluency Test Instrument Selection

In addition to the SAT 10 test, the research team was advised to include a measure of fluency at follow-up for the younger students in the reading sample. Younger students are more

likely to first show improvement in fluency before improving in overall comprehension, as measured by the SAT 10 standardized test (National Reading Panel 2000). Individually administered tests that are both short and fairly easy to administer were considered. The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) was selected and administered at follow-up to second- and third-graders in the reading centers, in addition to the SAT 10.

The DIBELS are "a set of standardized, individually administered measures of early literacy development. They are designed to be short (one minute) fluency measures used to monitor the development of pre-reading and early reading skills" (Dynamic Indicators of Basic Early Literacy Skills 2007a). DIBELS benchmark and progressive goals initially were derived based on data from all schools participating in the DIBELS Data System during the 2000-2001 and 2001-2002 academic years. And test-retest reliability for elementary students ranges from 0.92 to 0.97 (Dynamic Indicators of Basic Early Literacy Skills 2007a). Numerous additional studies have replicated the predictive utility of these goals in other, diverse samples. In this study, students were tested on measures of fluency — oral reading fluency (ORF) and nonsense word fluency (NWF).

The ORF assesses a child's skill in reading connected text. "Student performance is measured by having students read a passage aloud for one minute. Words omitted, substituted, and hesitations of more than three seconds are scored as errors. Words self-corrected within three seconds are scored as accurate. The number of correct words per minute from the passage is the oral reading fluency rate" (Dynamic Indicators of Basic Early Literacy Skills 2007b). Students in the study were asked to read three passages, and their median score was used in the analysis.

The NWF assesses a child's knowledge of "letter-sound correspondence and of the ability to blend letters into words in which letters represent their most common sounds (Dynamic Indicators of Basic Early Literacy Skills 2007c). The student is presented an 8.5-x-11-inch sheet of paper with randomly ordered vowel-consonant and consonant-vowel-consonant nonsense words (for example, sig, rav, ov) and is asked to produce verbally the individual letter-sound of each letter or to verbally produce, or read, the whole nonsense word. "For example, if the stimulus word is 'vaj,' the student could say /v/ /a/ /j/ or say the word /vaj/ to obtain a total of three letter-sounds correct. The student is allowed 1 minute to produce as many letter-sounds as he/she can, and the final score is the number of letter-sounds produced correctly in one minute. Because the measure is fluency based, students receive a higher score if they are phonologically recoding the word and receive a lower score if they are providing letter sounds in isolation" (Dynamic Indicators of Basic Early Literacy Skills 2007c).

School Record Data

The study also collected information about student performance on the locally administered tests from school record data and used these test scores as a supplementary measure of students' academic performance. The locally administered tests are also more likely to be a full battery and might measure math or reading more reliably than the abbreviated version of SAT 10 used by the study. On the other hand, these locally administered tests also may be testing a slightly different set of skills than tested by the abbreviated SAT 10. Thus, they provide a different measure of reading or math skill.

Each school district has its own specific test, so the closest measure to a total reading and total math score was used. (See Appendix Tables E.2 and E.3 for a list of math tests and reading tests available to the study sites.) In order to pool across the sites and estimate overall impact for the sample, each student's test score was standardized in the following way:

$$Z_{ij} = \frac{(Y_{ij} - \overline{Y}_j)}{s.d._j(Y_{ij})}$$

where:

Z_{ij} = the standardized score for student *i* from site *j*.

Y_{ij} = the raw score for student *i* from site *j* in the locally administered test.

$\overline{Y}_j$ = the average raw score for site *j* in the locally administered test.

$s.d._j(Y_{ij})$ = the standard deviation of the raw test scores for site *j*.

This transformed measure was then used as an outcome for student achievement.

Academic Behavior

Measures of students' academic behaviors come from the regular-school-day teacher survey conducted in the spring of the first program year. For each student in the study sample, the regular-school-day teacher was asked to fill out a short survey about any special academic support that the student receives during the school day and how the student behaved in the regular-school-day class. Specifically, teachers rated their students on the following:

Q6. How often does this student NOT complete homework?

Q7. How often is this student disruptive?

Q9. How often is this student attentive in class?

For each of these questions, the teacher was asked to choose from (1) Never, (2) Not very often, (3) Sometimes, and (4) Often. The answers, therefore, were coded on the scale of 1 to 4, with 1 indicating "Never" and 4 "Often."

The Evaluation of Academic Instruction in After-School Programs

Appendix Table E.2

Math District Tests, by State

Test	Criterion- or Norm-Referenced	Test Content
Standardized test administered to study students		
Stanford Achievement Test Series, 10th ed. (SAT 10) abbreviated battery	Norm-referenced	Number Sense and Operations; Patterns, Relationships, and Algebra; Geometry and Measurement; Data, Statistics, and Probability; Communication and Representation; Estimation; Mathematical Connections; Reasoning and Problem Solving; Mathematical Procedures
State-administered tests		
California Standards Tests (CST)	Criterion-referenced	Grade 3: Number Sense - Place Value, Addition and Subtraction; Number Sense - Multiplication, Division, and Fractions; Algebra and Functions; Measurement and Geometry; Statistics, Data Analysis, and Probability Grade 4: Number Sense - Decimals, Fractions, and Negative Numbers; Number Sense - Operations and Factoring; Algebra and Functions; Measurement and Geometry; Statistics, Data Analysis, and Probability Grade 5: Number Sense - Estimation, Percents, and Factoring; Number Sense - Operations with Fractions and Decimals; Algebra and Functions; Measurement and Geometry; Statistics, Data Analysis, and Probability
Connecticut Mastery Test (CMT)	Criterion-referenced	Numerical and Proportional Reasoning; Geometry and Measurement; Working with Data: Probability and Statistics; Algebraic Reasoning: Patterns and Functions; Integrated Understandings
Georgia Criterion Referenced Competency Tests (CRCT)	Criterion-referenced	Number Sense and Numeration; Geometry and Measurement; Patterns and Relationships; Statistics and Probability; Computation and Estimation; Problem Solving
Florida's Comprehensive Assessment Test (FCAT)	Criterion-referenced	Number Sense, Concepts, and Operations; Measurement; Geometry and Spatial Sense; Algebraic Thinking; Data Analysis and Probability

(continued)

Appendix Table E.2 (continued)

Test	Criterion- or Norm-Referenced	Test Content
Pennsylvania System of School Assessment (PSSA)	Criterion-referenced	Numbers and Operations; Measurement; Geometry; Algebraic Concepts; Data Analysis and Probability
Scantron Math (administered by the State of Kansas)	Norm-referenced	Algebra; Geometry; Measurement; Data Analysis & Probability; Number & Operations
Stanford Achievement Test Series, 10th ed. (SAT 10) full battery (administered by the State of Alabama)	Norm-referenced	Number Sense and Operations; Patterns, Relationships, and Algebra; Geometry and Measurement; Data, Statistics, and Probability; Communication and Representation; Estimation; Mathematical Connections: Reasoning and Problem Solving; Mathematical Procedures
Texas Assessment of Knowledge and Skills (TAKS)	Criterion-referenced	Numbers, Operations, and Quantitative Reasoning; Patterns, Relationships, and Algebraic Reasoning; Geometry and Spacial Reasoning Measurement; Probability and Statistics; Mathematical Processes and Tools
Wisconsin Knowledge and Concepts Examinations - Criterion Referenced Test (WKCE-CRT)	Criterion-referenced	Mathematical Process; Number Operations and Relationships; Geometry; Measurement; Statistics and Probability; Algebraic Relationships

SOURCES: Information on the Stanford Achievement Test Series, 10th ed. (SAT 10) abbreviated battery, was retrieved from the Harcourt Assessment Web site. State test names, formats, and contents were provided by in-house district data, test assessment Web sites, and state Department of Education Web sites.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table E.3

Reading District Tests, by State

Test	Criterion- or Norm-Referenced	Test Content
Standardized test administered to study students		
Stanford Achievement Test Series, 10th ed. (SAT 10) abbreviated battery	Norm-referenced	Reading Comprehension - initial understanding, interpretation, and critical analysis of reading passages; Reading Vocabulary - understanding of the printed word, synonyms, and multiple meaning words; Word Study Skills - structural and phonetic analysis, such as identifying and decoding compound words and contractions and recognizing sounds of consonants and vowels
State-administered tests		
California Standards Tests (CST)	Criterion-referenced	Word Analysis; Reading Comprehension; Literary Response and Analysis; Writing Strategies; Written Conventions
Georgia Criterion Referenced Competency Tests (CRCT)	Criterion-referenced	Vocabulary; Comprehension; Reading for Literacy Comprehension; Reading for Information; Reading Skills and Vocabulary Acquisiton; Functional and Media Literacy
Florida's Comprehensive Assessment Test (FCAT)	Criterion-referenced	Words and Phrases in Context; Main Idea, Plot, and Purpose; Comparisons and Cause/Effect; Reference and Research
Louisiana Educational Assessment Program (LEAP)	Criterion-referenced	Read, Comprehend, and Respond to a Range of Materials; Write Competently; Use Conventions of Language; Apply Speaking and Listening Skills [not assessed]; Locate, Select, and Synthesize Information; Read, Analyze, and Respond to Literature; Apply Reasoning and Problem-Solving Skills
New Mexico Standards Based Assessment (NMSBA)	Criterion-referenced	Reading and Listening for Comprehension; Writing and Speaking for Expression; Literature and Media

(continued)

Appendix Table E.3 (continued)

Test	Criterion- or Norm-Referenced	Test Content
New York State English Language Arts	Criterion-referenced	Understand Story Events; Draw Conclusions; Make Predictions; Identify the Main Idea; Use Text to Understand Unfamiliar Vocabulary Words; Identify Supporting Details; Identify Point of View; Evaluate Ideas Based on Prior Knowledge; Follow Ideas and Events in the Text; Distinguish Fact from Opinion; Understand Features That Distinguish Genres; Use Figurative Language to Interpret Text
Pennsylvania System of School Assessment (PSSA)	Criterion-referenced	Comprehension and Reading Skills; Interpretation and Analysis of Fiction and Non-Fiction Text
TerraNova (administered by the State of Michigan)	Norm-referenced	Basic Understanding; Analyze Text; Evaluate and Extend Meaning; Identify Reading Strategies
Texas Assessment of Knowledge and Skills (TAKS)	Criterion-referenced	Basic Understanding; Literary Elements; Analysis Using Reading Strategies; Analysis Using Critical-Thinking Skills
Wisconsin Knowledge and Concepts Examinations - Criterion Referenced Test (WKCE-CRT)	Criterion-referenced	Determine the Meaning of Words and Phrases in Context; Understand Text; Analyze Text; Evaluate and Extend Text

SOURCES: Information on the Stanford Achievement Test Series, 10th ed. (SAT 10) abbreviated battery, was retrieved from the Harcourt Assessment Web site. State test names, formats, and contents were provided by in-house district data, test assessment Web sites, and state Department of Education Web sites.

Appendix F

Estimating Effects and Assessing Robustness

This appendix provides a detailed discussion of the statistical model used to estimate the program impacts and other related statistical issues. It also discusses various tests that were used to assess the robustness of the impact estimates reported in the text and provides the results for these tests.

Analysis of Program Impacts

The program impact analysis involves examining outcome measures constructed from the follow-up student achievement tests, a survey of regular-school-day teachers, and student records from participating districts, with key outcomes listed in Chapter 2. Note that all the listed outcomes are measured at the level of individual students. These outcomes are used to calculate the estimates of impacts of each of the two academic programs separately (the math or the reading program) by comparing outcomes for the enhanced program group and the regular program group within the after-school centers and grade levels.

The Model

Impacts of reading and math programs were estimated separately. For each outcome, the basic model used in the analysis is the following:

$$Y_{ijk} = \gamma_0 Y_{-1,ijk} + \beta_0 T_{ijk} + \sum_k \sum_j \gamma_{1kj} B_{ijk} + \sum_S \gamma_{2s} X_{sijk} + \varepsilon_{jk} \qquad (1)$$

where:

T_{ijk} = one if student i from grade j in center k is assigned to the enhanced program and zero otherwise.

Y_{-1ijk} = the pretest score for student i from grade j in center k before random assignment.[1]

[1]Pretest scores are scaled scores from the SAT 10 (SAT 9 for a couple of centers) reading and math tests administered in the fall of 2005, before the start of the after-school program. Total scores for each subject were used in the analysis of respective samples.

B_{ijk} = block dummy, one if student *i* is in a particular random assignment block, defined by grade j, center *k,* and zero otherwise, $k = 1$ to 25, $j = 1$ to 4.

X_{sijk} = the *s* other student-level covariates for student *i* from grade *j* in center *k*.

ε_{ijk} = a student-level random error respectively, assumed to be independently and identically distributed.

The coefficient, β_0, represents the overall impact of being randomized to enhanced instruction instead of the regular after-school program *for an average student in the sample.* The traditional t-statistic for this coefficient tests whether the estimated average impact for the sample of students in the study centers is statistically significantly different from zero. This analysis does not attempt to generalize statistically beyond the observed sample of sites; thus, the traditional t-test is appropriate.

There are several features worth noting in this model:

- β_0 is a "fixed-effect" estimate that addresses the question: What is the program effect of enhanced instruction for the average student in the sample? This approach is taken because the goal of this study is to conduct an efficacy study of the effects of a new approach, and sites are not selected to be a random sample of a larger population of sites.
- Ordinary Least Squares (OLS) regression is used to estimate Equation (1).
- Indicators for each of the blocks used in the random assignment process (B_{ijk}, defined by the center and the given grade level of the student on the baseline questionnaire) are included in the model to reflect the design feature (that is, differential rates of treatment assignment, by block) and control for the variation in mean outcome level (which can be due to different characteristics of centers, school settings, and so on) across blocks.
- The model controls for individual-level pretest measure. This information can increase the precision of impact estimates, especially for fixed-effect models, because pretests substantially reduce random posttest error, which is the sole source of uncertainty in a fixed-effect model.
- Other baseline covariates are added to the model to improve precision. These covariates include student's gender, race/ethnicity, free/reduced-price lunch status, age, whether a student is from a single-adult household, whether a student is overage for grade, and the mother's education level.

The design also allows the research to detect effects among subgroups of students that are defined by characteristics depicting a student's pre-random assignment condition. To be parsimonious, subgroups on two theoretically relevant and policy-relevant characteristics were examined: subgroups based on students' grade levels and baseline academic performances.

Other Analytical Issues

Missing Covariates

For the baseline achievement test, there are 22 missing cases (11 for math and 11 for reading). For other covariates, there are very few (5 percent or less) missing cases.[2] To keep the sample as complete as possible, the missing values were imputed with the mean value of the center-by-grade-by-treatment-status block to which the student belongs.[3] If more than 5 percent of the observations are missing data for a given variable, then a dummy variable indicating whether a student is missing this covariate or not was also included.

Missing Outcome Measures

Missing data for outcomes pose a problem that is more serious and more difficult to solve because it requires omitting sample members from the impact analysis, which can produce selection bias if this attrition is substantial and nonrandom. As discussed in Appendix C, response rates in this study were in general above 85 percent, and the student characteristics of the full study sample and the analysis sample are similar. Therefore, of the full sample, 147 math (7 percent) and 235 reading (11 percent) students with missing outcome measures were excluded from the impact analysis sample.

[2]Among the students in the reading sites, 4 are missing a race/ethnicity indicator; 83 are missing a free lunch status indicator; 19 are missing information about single-adult household; and 92 are missing information about mother's education. Among the students in the math sites, 2 are missing a race/ethnicity indicator; 58 are missing a free lunch status indicator; 36 are missing information about single-adult household; and 121 are missing information about mother's education. (No students are missing indicators of gender or age.)

[3]Rather than imputing the missing reading or math SAT-10 total scaled score, the mean score for the missing subtest raw score was imputed, and then the subtest raw scores were added, and that student was assigned a scaled score for the given raw score. Thus, if there is an actual score for one or more of the subtests, the imputed total score will incorporate the actual subtest scores.

Additional Tests and Checks

For the Math Sample

In addition to the math program impact results presented in Chapter 4 of the report, the program's impacts on student performance in locally administered math tests were also estimated, to compare with those on SAT 10 tests. The locally administered tests are mostly full-battery tests and might measure math skills more reliably than the abbreviated tests used by the study.

An important caveat for this comparison relates to data availability. The locally administered test data were not always available for second-graders in those study sites that start testing students in the third grade. As a result, all second-graders were excluded from this analysis, and the total sample size for the locally administered test analysis is 1,310 for math.

Appendix Table F.1 presents the estimated program impacts on student performance in locally administered tests for math. Because these test scores were standardized within each study site, all estimated impacts are in effect size units.[4] The table also shows the program impact on the study-administered SAT 10 tests for the sample of students whose local test scores were available for comparison purpose. Because second-graders were excluded from the analysis, the table does not show impact estimates for total scores for the subgroup of second- and third-graders.

For the math sample, all five estimates have the same sign in both measures. The estimated effect sizes for the local tests are in the same direction but with differing magnitudes than those estimated for study-administered SAT 10 total scores, and they are not statistically significant. On the other hand, the program impacts on the SAT 10 math total scores are statistically significant for the subgroup of fourth- and fifth-graders and for the subgroup of students who performed at "basic" level before the program started. This pattern for subgroup findings is the same as the one shown in Table 4.1 for the math analysis sample. Furthermore, the following checks were conducted to see whether the impact estimates on SAT 10 test scores are robust:

- All impacts were reestimated for the sample of all SAT 10 respondents to make sure that no imbalance was created when the full study sample was limited to the analysis sample.

[4]Appendix E describes the standardization of the test score variable.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table F.1

Impact of the Enhanced Math Program on Student Achievement for Grades 3 to 5

Student Achievement Outcome	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
State test analysis sample					
State test scaled scores	0.02	-0.01	0.03	0.03	0.49
SAT 10 math total scaled scores	620.62	618.35	2.26	0.05	0.08
Sample size (total = 1,310)	729	581			
Grade subgroup					
Grades 4 and 5					
State test scaled scores	0.03	-0.06	0.09	0.09	0.08
SAT 10 math total scaled scores	626.33	622.27	4.06 *	0.09	0.01
Sample size (total = 921)	515	406			
Prior-achievement subgroups					
Students scoring at below basic level					
State test scaled scores	-0.72	-0.77	0.06	0.06	0.54
SAT 10 math total scaled scores	594.09	591.54	2.56	0.06	0.34
Sample size (total = 347)	184	163			
Students scoring at basic level					
State test scaled scores	0.07	0.02	0.05	0.05	0.40
SAT 10 math total scaled scores	618.83	614.98	3.86 *	0.09	0.04
Sample size (total = 679)	397	282			
Students scoring at proficient level					
State test scaled scores	0.72	0.83	-0.11	-0.11	0.39
SAT 10 math total scaled scores	653.78	656.57	-2.79	-0.06	0.46
Sample size (total = 239)	126	113			

(continued)

SOURCES: MDRC calculations are from results on state tests administered in the 2005-2006 school year and follow-up results on the Stanford Achievement Test Series, 10th ed. (SAT 10) abbreviated battery.

NOTES: State test data were not available for most second-graders because many of the study sites begin testing students in the third grade, and, as a result, all second-graders are excluded from this analysis. In addition, the state test analysis sample is restricted to those from the full analysis sample for whom a state test score was obtained. The resulting state test analysis sample represents 88 percent of the third- through fifth-graders in the full analysis sample and is used to calculate the SAT 10 and state test findings presented.

Each student's state test score was converted into a standardized score because school districts in different states administer different tests. See Appendix E for details.

Appendix Table F.1 (continued)

Based on the SAT 10 national norming sample, math total scaled scores have the following possible ranges: for the state test analysis sample, scores range from 428 to 796; for the fourth- and fifth-grade subgroup, scores range from 450 to 796.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for indicators of random assignment, baseline math total scaled score, race/ethnicity, gender, free-lunch status, age, overage for grade, single-adult household, and mother's education. The values in column 1 (labeled "Enhanced Program") are the observed mean for the members randomly assigned to the enhanced program group. The regular program group values in column 2 are the regression-adjusted means using the observed mean covariate values for the enhanced program group as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each impact estimate. Statistical significance is indicated by (*) when the p-value is less than or equal to 5 percent.

The estimated impact effect size of the state test score is calculated as a proportion of the state test score standard deviation of the regular program group from the state test analysis sample. The estimated impact effect size of the SAT 10 math total scaled score is calculated as a proportion of the standard deviation of the regular program group from the full analysis sample, which is 44.64. The standard deviation of a SAT 10 national norming sample with the same grade composition as the full analysis sample is 39.00.

There are 22 enhanced program group students and 23 regular program group students who performed at the advanced level on the baseline SAT 10; they are excluded from the prior-achievement subgroup analysis.

This change in the sample added 19 observations for the math sample. Appendix Table F.2 presents student achievement impact results for math, using the SAT 10 respondents from the full study sample. The general patterns of the findings do not change at all.

- All impacts were reestimated with a model that has no covariates other than the "block" (random assignment unit) indicators, the treatment status indicator, and prior achievement.

In other words, the following model was used to estimate the program impacts:

$$Y_{ijk} = \sum_{m}\sum_{n} \gamma_{0mn} B_{ijk} + \beta_0 T_{ijk} + \gamma_1 Y_{-1,ijk} + \varepsilon_{ijk} \qquad (2)$$

The variables are defined as before. Because this study is based on a randomized experiment, both sets of estimates — those with or those without controlling for other baseline characteristics — provide an unbiased estimate of the treatment effect. The precision of the estimated impact, however, is likely improved by controlling for other baseline characteristics.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table F.2

Impact of the Enhanced Math Program on Student Achievement for the SAT 10 Respondent Sample

Student Achievement Outcome	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
SAT 10 respondent sample					
SAT 10 math total scaled scores	604.73	601.99	2.73 *	0.06	0.01
Problem solving	605.85	603.40	2.45 *	0.05	0.04
Procedures	604.91	600.81	4.11 *	0.08	0.01
Sample size (total = 1,980)	1,093	887			
Grade subgroups					
Grades 2 and 3					
SAT 10 math total scaled scores	582.83	581.07	1.76	0.04	0.29
Problem solving	584.43	583.60	0.83	0.02	0.62
Procedures	583.21	579.11	4.10	0.08	0.08
Sample size (total = 984)	542	442			
Grades 4 and 5					
SAT 10 math total scaled scores	626.26	622.55	3.71 *	0.08	0.01
Problem solving	626.88	622.76	4.12 *	0.09	0.01
Procedures	626.26	622.25	4.01 *	0.07	0.05
Sample size (total = 996)	551	445			
Prior-achievement subgroups					
Students scoring at below basic level					
SAT 10 math total scaled scores	583.67	580.85	2.82	0.06	0.22
Problem solving	585.91	582.93	2.98	0.07	0.23
Procedures	579.20	576.70	2.50	0.05	0.43
Sample size (total = 474)	243	231			
Students scoring at basic level					
SAT 10 math total scaled scores	600.28	597.00	3.28 *	0.07	0.03
Problem solving	601.53	598.03	3.50 *	0.08	0.03
Procedures	600.48	595.58	4.89 *	0.09	0.02
Sample size (total = 1,062)	616	446			
Students scoring at proficient level					
SAT 10 math total scaled scores	634.24	631.27	2.98	0.07	0.31
Problem solving	633.61	629.98	3.63	0.08	0.22
Procedures	639.74	637.54	2.20	0.04	0.61
Sample size (total = 384)	205	179			

(continued)

Appendix Table F.2 (continued)

SOURCE: MDRC calculations are from follow-up results on the Stanford Achievement Test Series, 10th ed. (SAT 10) abbreviated battery.

NOTES: The SAT 10 respondent sample is composed of all students from the full study sample who have a follow-up SAT 10 math total score.

Based on the SAT 10 national norming sample, total, problem solving, and procedures scaled scores, respectively, have the following possible ranges: for the SAT 10 respondent sample, scores range from 389 to 796, 414 to 776, and 413 to 768; for the second- and third-grade subgroup, scores range from 389 to 741, 414 to 719, and 413 to 715; and for the fourth- and fifth-grade subgroup, scores range from 450 to 796, 468 to 776, and 485 to 768.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for indicators of random assignment, baseline math total scaled score, race/ethnicity, gender, free-lunch status, age, overage for grade, single-adult household, and mother's education. The values in column 1 (labeled "Enhanced Program") are the observed mean for the members randomly assigned to the enhanced program group. The regular program group values in column 2 are the regression-adjusted means using the observed mean covariate values for the enhanced program group as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each impact estimate. Statistical significance is indicated by (*) when the p-value is less than or equal to 5 percent.

The estimated impact effect size of the SAT 10 math total scaled score is calculated as a proportion of the standard deviation of the regular program group from the analysis sample, which is 44.64. The standard deviation of a SAT 10 national norming sample with the same grade composition as the study sample is 39.00. For each subtest, the estimated impact effect size is calculated as a proportion of the standard deviation of the regular program group from the analysis sample.

There are 29 enhanced program group students and 31 regular program group students who performed at the advanced level on the baseline SAT 10; they are excluded from the prior-achievement subgroup analysis.

As can be seen from Appendix Table F.3, dropping these covariates from the model affected the precision of the impact estimates but did not affect the magnitudes or the patterns of the impact findings, as one would expect from a randomized experiment.

- All impacts were reestimated with a model that has no covariates other than the "block" (random assignment unit) indicators and the treatment status indicator.

In other words, the following model was used to estimate the program impacts:

$$Y_{ijk} = \sum_{m}\sum_{n} \gamma_{0mn} B_{ijk} + \beta_0 T_{ijk} + \varepsilon_{ijk} \qquad (3)$$

The variables are defined as before and, as can be seen from Appendix Table F.4, dropping covariates from the model and controlling only for the randomization strata did not affect the magnitudes of the impact findings, but statistical significance levels differ in some cases due to less statistical precision.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table F.3

Impact of the Enhanced Math Program on Student Achievement for the Analysis Sample Without Demographic Characteristics as Model Covariates

Student Achievement Outcome	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
Analysis sample					
SAT 10 math total scaled scores	605.10	602.16	2.94 *	0.07	0.01
Problem solving	606.15	603.70	2.45 *	0.05	0.04
Procedures	605.30	600.74	4.56 *	0.08	0.00
Sample size (total = 1,961)	1,081	880			
Grade subgroups					
Grades 2 and 3					
SAT 10 math total scaled scores	583.23	581.13	2.10	0.05	0.21
Problem solving	584.82	583.77	1.04	0.02	0.54
Procedures	583.55	578.93	4.62	0.09	0.05
Sample size (total = 971)	533	438			
Grades 4 and 5					
SAT 10 math total scaled scores	626.37	622.63	3.74 *	0.08	0.01
Problem solving	626.91	623.08	3.83 *	0.09	0.02
Procedures	626.46	622.02	4.45 *	0.08	0.03
Sample size (total = 990)	548	442			
Prior-achievement subgroups					
Students scoring at below basic level					
SAT 10 math total scaled scores	584.29	581.92	2.37	0.05	0.30
Problem solving	586.30	583.81	2.49	0.06	0.31
Procedures	580.17	578.19	1.99	0.04	0.53
Sample size (total = 467)	239	228			
Students scoring at basic level					
SAT 10 math total scaled scores	600.52	597.29	3.24 *	0.07	0.03
Problem solving	601.74	598.29	3.45 *	0.08	0.04
Procedures	600.63	595.80	4.83 *	0.09	0.02
Sample size (total = 1,055)	612	443			
Students scoring at proficient level					
SAT 10 math total scaled scores	634.67	632.03	2.64	0.06	0.36
Problem solving	634.02	631.71	2.32	0.05	0.43
Procedures	640.08	637.08	3.00	0.06	0.48
Sample size (total = 380)	202	178			

(continued)

Appendix Table F.3 (continued)

SOURCE: MDRC calculations are from follow-up results on the Stanford Achievement Test Series, 10th ed. (SAT 10) abbreviated battery.

NOTES: Based on the SAT 10 national norming sample, total, problem solving, and procedures scaled scores, respectively, have the following possible ranges: for the analysis sample, scores range from 389 to 796, 414 to 776, and 413 to 768; for the second- and third-grade subgroup, scores range from 389 to 741, 414 to 719, and 413 to 715; and for the fourth- and fifth-grade subgroup, scores range from 450 to 796, 468 to 776, and 485 to 768.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for indicators of random assignment and baseline math total scaled score. The values in column 1 (labeled "Enhanced Program") are the observed mean for the members randomly assigned to the enhanced program group. The regular program group values in column 2 are the regression-adjusted means using the observed mean covariate values for the enhanced program group as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each impact estimate. Statistical significance is indicated by (*) when the p-value is less than or equal to 5 percent.

The estimated impact effect size of the SAT 10 math total scaled score is calculated as a proportion of the standard deviation of the regular program group, which is 44.64 based on the analysis sample and the model controlling for demographic characteristics. The standard deviation of a SAT 10 national norming sample with the same grade composition as the study sample is 39.00. For each subtest, the estimated impact effect size is calculated as a proportion of the standard deviation of the regular program group.

There are 28 enhanced program group students and 31 regular program group students who performed at the advanced level on the baseline SAT 10; they are excluded from the prior-achievement subgroup analysis.

In summary, the program impacts on the locally administered math test have the same sign as the study-administered SAT 10 impacts but are not statistically significant. The two robustness checks demonstrated that the math impact results reported in Chapter 4 are not affected by the various sample restriction and the alternative model specifications.

For the Reading Sample

Similar to the analysis for the math program, the program impacts on student performance in locally administered reading tests were estimated to compare with those on SAT 10 reading tests.

The locally administered test data were not available for all second-graders in those study sites that start testing students in the third grade. As a result, all second-graders were excluded from this analysis, and the total sample size for the locally administered test analysis is 1,238 for reading.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table F.4

Impact of the Enhanced Math Program on Student Achievement for the Analysis Sample With a Random Assignment Indicator as the Only Model Covariate

Student Achievement Outcome	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
Analysis sample					
SAT 10 math total scaled scores	605.10	601.99	3.11 *	0.07	0.03
Problem solving	606.15	603.53	2.62	0.06	0.08
Procedures	605.30	600.55	4.76 *	0.09	0.01
Sample size (total = 1,961)	1,081	880			
Grade subgroups					
Grades 2 and 3					
SAT 10 math total scaled scores	583.23	580.61	2.62	0.06	0.23
Problem solving	584.82	583.29	1.53	0.03	0.47
Procedures	583.55	578.31	5.24	0.10	0.07
Sample size (total = 971)	533	438			
Grades 4 and 5					
SAT 10 math total scaled scores	626.37	622.78	3.59	0.08	0.07
Problem solving	626.91	623.22	3.69	0.08	0.08
Procedures	626.46	622.17	4.29	0.08	0.09
Sample size (total = 990)	548	442			
Prior-achievement subgroups					
Students scoring at below basic level					
SAT 10 math total scaled scores	584.29	580.72	3.57	0.08	0.14
Problem solving	586.30	582.48	3.82	0.09	0.15
Procedures	580.17	577.11	3.06	0.06	0.35
Sample size (total = 467)	239	228			
Students scoring at basic level					
SAT 10 math total scaled scores	600.52	597.41	3.11	0.07	0.06
Problem solving	601.74	598.40	3.33	0.07	0.06
Procedures	600.63	595.95	4.68 *	0.09	0.04
Sample size (total = 1,055)	612	443			
Students scoring at proficient level					
SAT 10 math total scaled scores	634.67	632.16	2.50	0.06	0.41
Problem solving	634.02	631.84	2.19	0.05	0.47
Procedures	640.08	637.23	2.85	0.05	0.52
Sample size (total = 380)	202	178			

(continued)

Appendix Table F.4 (continued)

SOURCE: MDRC calculations are from follow-up results on the Stanford Achievement Test Series, 10th ed. (SAT 10) abbreviated battery.

NOTES: Based on the SAT 10 national norming sample, total, problem solving, and procedures scaled scores, respectively, have the following possible ranges: for the analysis sample, scores range from 389 to 796, 414 to 776, and 413 to 768; for the second- and third-grade subgroup, scores range from 389 to 741, 414 to 719, and 413 to 715; and for the fourth- and fifth-grade subgroup, scores range from 450 to 796, 468 to 776, and 485 to 768.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for indicators of random assignment strata. The values in column 1 (labeled "Enhanced Program") are the observed mean for the members randomly assigned to the enhanced program group. The regular program group values in column 2 are the regression-adjusted means using the observed distribution of the enhanced program group across random assignment strata as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each impact estimate. Statistical significance is indicated by (*) when the p-value is less than or equal to 5 percent.

The estimated impact effect size of the SAT 10 math total scaled score is calculated as a proportion of the standard deviation of the regular program group, which is 44.64 based on the analysis sample and the model controlling for demographic characteristics. The standard deviation of a SAT 10 national norming sample with the same grade composition as the study sample is 39.00. For each subtest, the estimated impact effect size is calculated as a proportion of the standard deviation of the regular program group.

There are 28 enhanced program group students and 31 regular program group students who performed at the advanced level on the baseline SAT 10; they are excluded from the prior-achievement subgroup analysis.

Appendix Table F.5 presents the estimated program impacts on student performance in locally administered tests for reading. Because these test scores were standardized within each study site, all estimated impacts are in effect size.[5] The table also shows the program impact on the study-administered SAT 10 tests for the sample of students whose local test scores were available for comparison purpose. Because second-graders were excluded from the analysis, the tables does not show impact estimates for total scores for the subgroup of second- and third-graders.

For the reading sample, the estimated impact effect size using the local test is –0.01, and that using the study-administered SAT 10 total test score is –0.01 too. None of these estimates are statistically different from zero. Overall, the locally administered tests do not yield qualitatively different findings about the program impact.

In addition, the following checks were conducted to see whether the estimated reading program impacts reported in Chapter 6 are robust:

[5]Appendix E describes the standardization of the test score variable.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table F.5

Impact of the Enhanced Reading Program on Student Achievement for Grades 3 to 5

Student Achievement Outcome	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
State test analysis sample					
State test scaled scores	-0.04	-0.03	-0.01	-0.01	0.92
SAT 10 reading total scaled scores	598.74	599.07	-0.33	-0.01	0.77
Sample size (total = 1,238)	720	518			
Grade subgroup					
Grades 4 and 5					
State test scaled scores	-0.05	-0.03	-0.01	-0.01	0.81
SAT 10 reading total scaled scores	605.38	605.71	-0.33	-0.01	0.81
Sample size (total = 830)	486	344			
Prior-achievement subgroups					
Students scoring at below basic level					
State test scaled scores	-0.40	-0.40	0.00	0.01	0.95
SAT 10 reading total scaled scores	585.68	583.33	2.35	0.07	0.16
Sample size (total = 564)	342	222			
Students scoring at basic level					
State test scaled scores	0.20	0.24	-0.04	-0.04	0.56
SAT 10 reading total scaled scores	606.61	608.85	-2.24	-0.06	0.19
Sample size (total = 580)	335	245			
Students scoring at proficient level					
State test scaled scores	0.97	0.70	0.27	0.28	0.26
SAT 10 reading total scaled scores	640.22	632.81	7.41	0.21	0.42
Sample size (total = 90)	41	49			

(continued)

SOURCES: MDRC calculations are from results on state tests administered in the 2005-2006 school year and follow-up results on the Stanford Achievement Test Series, 10th ed. (SAT 10) abbreviated battery.

NOTES: State test data were not available for most second-graders because many of the study sites begin testing students in the third grade, and, as a result, all second-graders are excluded from this analysis. In addition, the state test analysis sample is restricted to those from the full analysis sample for whom a state test score was obtained. The resulting state test analysis sample represents 90 percent of the third- through fifth-graders in the full analysis sample and is used to calculate the SAT 10 and state test findings presented.

Appendix Table F.5 (continued)

Each student's test score was converted into a standardized score because school districts in different states administer different tests. See Appendix E for details.

Based on the SAT 10 national norming sample, reading total scaled scores have the following possible ranges: for the state test analysis sample, scores range from 416 to 787; for the fourth- and fifth-grade subgroup, scores range from 434 to 787.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for indicators of random assignment, baseline reading total scaled score, race/ethnicity, gender, free-lunch status, age, overage for grade, single-adult household, and mother's education. The values in column 1 (labeled "Enhanced Program") are the observed mean for the members randomly assigned to the enhanced program group. The regular program group values in column 2 are the regression-adjusted means using the observed mean covariate values for the enhanced program group as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each impact estimate. Statistical significance is indicated by (*) when the p-value is less than or equal to 5 percent.

The estimated impact effect size of the state test score is calculated as a proportion of the state test score standard deviation of the regular program group from the state test analysis sample. The estimated impact effect size of the SAT 10 reading total scaled score is calculated as a proportion of the standard deviation of the regular program group from the full analysis sample, which is 35.71. The standard deviation of a SAT 10 national norming sample with the same grade composition as the full analysis sample is 39.05.

There are 2 enhanced program group students and 2 regular program group students who performed at the advanced level on the baseline SAT 10; they are excluded from the prior-achievement subgroup analysis.

- All impacts were reestimated for the sample of all SAT 10 respondents to make sure that no imbalance was created when the full study sample was limited to the analysis sample.

This change in the sample added 76 observations for the reading sample. Appendix Table F.6 presents student achievement impact results for reading using the SAT 10 respondents from the full study sample. The general patterns of the findings do not change at all.

- All impacts were reestimated with a model that has no covariates other than the "block" (random assignment unit) indicators, the treatment status indicator, and prior achievement.

The model used here is the same as Equation (2). As can be seen from Appendix Table F.7, dropping these covariates from the model affected the significance level of the impact estimates but did not affect the magnitudes or the patterns of the impact findings, as one would expect from a randomized experiment.

- All impacts were reestimated with a model that has no covariates other than the "block" (random assignment unit) indicators and the treatment status indicator.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table F.6

Impact of the Enhanced Reading Program on Student Achievement for the SAT 10 Respondent Sample

Student Achievement Outcome	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
SAT 10 respondent sample					
SAT 10 reading total scaled scores	587.33	587.75	-0.42	-0.01	0.64
Vocabulary	580.74	580.39	0.35	0.01	0.79
Reading comprehension	588.51	589.03	-0.52	-0.01	0.66
Word study skills (grades 2-4)[a]	586.52	588.25	-1.73	-0.05	0.29
Sample size (total = 1,904)	1,092	812			
Grade subgroups					
Grades 2 and 3					
SAT 10 reading total scaled scores	569.10	569.73	-0.63	-0.02	0.63
Vocabulary	556.70	556.69	0.00	0.00	1.00
Reading comprehension	571.02	571.25	-0.23	-0.01	0.90
Word study skills	579.12	582.20	-3.08	-0.08	0.10
Sample size (total = 944)	544	400			
Grades 4 and 5					
SAT 10 reading total scaled scores	605.43	605.48	-0.05	0.00	0.97
Vocabulary	604.69	603.77	0.92	0.02	0.60
Reading comprehension	605.90	606.36	-0.46	-0.01	0.77
Sample size (total = 960)	548	412			
Prior-achievement subgroups					
Students scoring at below basic level					
SAT 10 reading total scaled scores	576.60	574.94	1.67	0.05	0.23
Vocabulary	567.71	565.92	1.78	0.04	0.38
Reading comprehension	579.00	576.54	2.46	0.06	0.17
Word study skills[a]	571.58	572.40	-0.82	-0.02	0.75
Sample size (total = 770)	456	314			
Students scoring at basic level					
SAT 10 reading total scaled scores	591.74	593.37	-1.63	-0.05	0.24
Vocabulary	586.18	587.34	-1.16	-0.03	0.57
Reading comprehension	591.81	593.62	-1.80	-0.05	0.31
Word study skills[a]	591.38	595.05	-3.67	-0.10	0.12
Sample size (total = 912)	521	391			

(continued))

Appendix Table F.6 (continued)

Student Achievement Outcome	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
Students scoring at proficient level					
SAT 10 reading total scaled scores	608.36	610.83	-2.47	-0.07	0.54
Vocabulary	605.58	605.76	-0.17	0.00	0.97
Reading comprehension	609.28	612.77	-3.50	-0.09	0.51
Word study skills[a]	612.02	610.62	1.40	0.04	0.82
Sample size (total = 207)	107	100			

SOURCE: MDRC calculations are from follow-up results on the Stanford Achievement Test Series, 10th ed. (SAT 10) abbreviated battery.

NOTES: The SAT 10 respondent sample is composed of all students from the full study sample who have a follow-up SAT 10 reading total score.

Based on the SAT 10 national norming sample, total, reading comprehension, vocabulary, and word study skills scaled scores, respectively, have the following possible ranges: for the SAT 10 respondent sample, scores range from 374 to 787, 439 to 777, 412 to 739, and 410 to 740; for the second- and third-grade subgroup, scores range from 374 to 765, 439 to 743, 412 to 700, and 410 to 727; and for the fourth- and fifth-grade subgroup, scores range from 434 to 787, 478 to 777, and 484 to 739.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for indicators of random assignment, baseline reading total scaled score, race/ethnicity, gender, free-lunch status, age, overage for grade, single-adult household, and mother's education. The values in column 1 (labeled "Enhanced Program") are the observed mean for the members randomly assigned to the enhanced program group. The regular program group values in column 2 are the regression-adjusted means using the observed mean covariate values for the enhanced program group as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each impact estimate. Statistical significance is indicated by (*) when the p-value is less than or equal to 5 percent.

The estimated impact effect size of the SAT 10 reading total scaled score is calculated as a proportion of the standard deviation of the regular program group from the analysis sample, which is 35.71. The standard deviation of a SAT 10 national norming sample with the same grade composition as the study sample is 39.05. For each subtest, the estimated impact effect size is calculated as a proportion of the standard deviation of the regular program group from the analysis sample.

There are 8 enhanced program group students and 7 regular program group students who performed at the advanced level on the baseline SAT 10; they are excluded from the prior-achievement subgroup analysis.

[a]The sample consists of second- through fourth-graders only because the spring administration of the test to fifth-graders does not include word study skills.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table F.7

Impact of the Enhanced Reading Program on Student Achievement for the Analysis Sample Without Demographic Characteristics as Model Covariates

Student Achievement Outcome	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
Analysis sample					
SAT 10 reading total scaled scores	587.42	588.23	-0.81	-0.02	0.39
Vocabulary	580.94	580.91	0.03	0.00	0.98
Reading comprehension	588.72	589.54	-0.82	-0.02	0.50
Word study skills (grades 2-4)[a]	586.39	588.47	-2.08	-0.05	0.21
Sample size (total = 1,828)	1,048	780			
Grade subgroups					
Grades 2 and 3					
SAT 10 reading total scaled scores	569.42	570.61	-1.19	-0.03	0.38
Vocabulary	557.05	558.08	-1.03	-0.02	0.61
Reading comprehension	571.54	571.88	-0.33	-0.01	0.85
Word study skills	579.28	582.85	-3.57	-0.09	0.06
Sample size (total = 912)	524	388			
Grades 4 and 5					
SAT 10 reading total scaled scores	605.43	605.85	-0.42	-0.01	0.75
Vocabulary	604.84	603.69	1.15	0.02	0.52
Reading comprehension	605.89	607.11	-1.23	-0.03	0.44
Sample size (total = 916)	524	392			
Prior-achievement subgroups					
Students scoring at below basic level					
SAT 10 reading total scaled scores	577.48	575.57	1.91	0.05	0.19
Vocabulary	568.88	566.39	2.48	0.05	0.24
Reading comprehension	579.82	577.50	2.33	0.06	0.21
Word study skills[a]	572.06	571.54	0.52	0.01	0.84
Sample size (total = 736)	437	299			
Students scoring at basic level					
SAT 10 reading total scaled scores	591.61	593.71	-2.10	-0.06	0.13
Vocabulary	585.88	587.93	-2.05	-0.04	0.31
Reading comprehension	592.00	593.94	-1.94	-0.05	0.28
Word study skills[a]	591.06	595.26	-4.20	-0.11	0.08
Sample size (total = 877)	501	376			

(continued)

Appendix Table F.7 (continued)

Student Achievement Outcome	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
Students scoring at proficient level					
SAT 10 reading total scaled scores	606.71	612.62	-5.91	-0.17	0.13
Vocabulary	604.77	606.11	-1.34	-0.03	0.81
Reading comprehension	607.70	615.35	-7.65	-0.20	0.15
Word study skills[a]	610.92	612.44	-1.52	-0.04	0.81
Sample size (total = 201)	103	98			

SOURCE: MDRC calculations are from follow-up results on the Stanford Achievement Test Series, 10th ed. (SAT 10) abbreviated battery.

NOTES: Based on the SAT 10 national norming sample, total, reading comprehension, vocabulary, and word study skills scaled scores, respectively, have the following possible ranges: for the analysis sample, scores range from 374 to 787, 439 to 777, 412 to 739, and 410 to 740; for the second- and third-grade subgroup, scores range from 374 to 765, 439 to 743, 412 to 700, and 410 to 727; and for the fourth- and fifth-grade subgroup, scores range from 434 to 787, 478 to 777, and 484 to 739.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for indicators of random assignment and baseline reading total scaled score. The values in column 1 (labeled "Enhanced Program") are the observed mean for the members randomly assigned to the enhanced program group. The regular program group values in column 2 are the regression-adjusted means using the observed mean covariate values for the enhanced program group as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each impact estimate. Statistical significance is indicated by (*) when the pvalue is less than or equal to 5 percent.

The estimated impact effect size of the SAT 10 reading total scaled score is calculated as a proportion of the standard deviation of the regular program group, which is 35.71 based on the analysis sample and the model controlling for demographic characteristics. The standard deviation of a SAT 10 national norming sample with the same grade composition as the study sample is 39.05. For each subtest, the estimated impact effect size is calculated as a proportion of the standard deviation of the regular program group.

There are 7 enhanced program group students and 7 regular program group students who performed at the advanced level on the baseline SAT 10; they are excluded from the prior-achievement subgroup analysis.

[a]The sample consists of second- through fourth-graders only because the spring administration of the test to fifth-graders does not include word study skills.

The model used here is the same as Equation (3) and, as can be seen from Appendix Table F.8, dropping covariates from the model and controlling only for the randomization strata affected the precision of the impact estimates as well as the magnitudes and the patterns of the impact findings. This is because there were significant differences between the enhanced reading program group and the regular reading program group at baseline, which are no longer being controlled for in this model.

- The baseline reading scores of the regular program group and the enhanced program group were statistically different from each other in a number of reading sample blocks. After restricting the sample to those blocks where the baseline scores were similar, all impacts were then reestimated. (This restricted sample is 87 percent of the analysis sample.)

Even with randomization there may be differences in baseline characteristics between the enhanced and regular program groups that are attributable to chance. Recall from Chapter 5 that there were statistically significant differences between the enhanced reading group and the regular reading program group at baseline. As a robustness check, block-by-block baseline differences in test scores were checked, and 12 blocks with the biggest baseline test score differences were excluded from the sample.[6] The remaining sample achieved balance between the enhanced program group and the regular program group at baseline. All impacts were reestimated using this restricted sample. This series of tests yields very similar impact estimates for the reading program sample (see Appendix Table F.9). These results show that controlling for the baseline characteristics as covariates in the impact model sufficiently eliminated the observed baseline differences between the enhanced program group and the regular program group.

In general, the reading impact results reported in Chapter 6 of this report are not affected by the various sample restriction and alternative model specifications.[7]

[6]A block was dropped if the baseline total reading test score difference between the enhanced program and regular program groups within that block was bigger than the overall difference between these groups by more than 1.75 standard deviations.

[7]In addition, 15 percent of parents reported on applications that the primary language spoken at home is Spanish. Since the classes were taught in English, one concern was that students who primarily do not speak English were not able to benefit from the program. Impacts were reestimated for those students who did not indicate that Spanish is the primary language spoken at home, and the results did not change.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table F.8

mpact of the Enhanced Reading Program on Student Achievement for the Analysi Sample With a Random Assignment Indicator as the Only Model Covariate

Student Achievement Outcome	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
Analysis sample					
SAT 10 reading total scaled scores	587.42	590.68	-3.26 *	-0.09	0.01
Vocabulary	580.94	583.98	-3.04	-0.07	0.08
Reading comprehension	588.72	591.84	-3.12 *	-0.08	0.03
Word study skills (grades 2-4)[a]	586.39	591.30	-4.91 *	-0.13	0.01
Sample size (total = 1,828)	1,048	780			
Grade subgroups					
Grades 2 and 3					
SAT 10 reading total scaled scores	569.42	574.02	-4.60 *	-0.13	0.02
Vocabulary	557.05	562.47	-5.43 *	-0.12	0.04
Reading comprehension	571.54	575.30	-3.76	-0.10	0.10
Word study skills	579.28	585.67	-6.39 *	-0.17	0.00
Sample size (total = 912)	524	388			
Grades 4 and 5					
SAT 10 reading total scaled scores	605.43	607.34	-1.91	-0.05	0.26
Vocabulary	604.84	605.48	-0.65	-0.01	0.77
Reading comprehension	605.89	608.38	-2.49	-0.06	0.18
Sample size (total = 916)	524	392			
Prior-achievement subgroups					
Students scoring at below basic level					
SAT 10 reading total scaled scores	577.48	576.35	1.13	0.03	0.46
Vocabulary	568.88	567.49	1.39	0.03	0.53
Reading comprehension	579.82	578.15	1.67	0.04	0.38
Word study skills[a]	572.06	571.94	0.12	0.00	0.96
Sample size (total = 736)	437	299			
Students scoring at basic level					
SAT 10 reading total scaled scores	591.61	594.75	-3.14 *	-0.09	0.04
Vocabulary	585.88	589.24	-3.35	-0.07	0.12
Reading comprehension	592.00	594.87	-2.87	-0.07	0.12
Word study skills[a]	591.06	596.46	-5.40 *	-0.14	0.03
Sample size (total = 877)	501	376			

(continued)

Appendix Table F.8 (continued)

Student Achievement Outcome	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
Students scoring at proficient level					
SAT 10 reading total scaled scores	606.71	613.07	-6.36	-0.18	0.12
Vocabulary	604.77	606.85	-2.09	-0.05	0.73
Reading comprehension	607.70	615.75	-8.05	-0.21	0.14
Word study skills[a]	610.92	612.35	-1.44	-0.04	0.82
Sample size (total = 201)	103	98			

SOURCE: MDRC calculations are from follow-up results on the Stanford Achievement Test Series, 10th ed. (SAT 10) abbreviated battery.

NOTES: Based on the SAT 10 national norming sample, total, reading comprehension, vocabulary, and word study skills scaled scores, respectively, have the following possible ranges: for the analysis sample, scores range from 374 to 787, 439 to 777, 412 to 739, and 410 to 740; for the second- and third-grade subgroup, scores range from 374 to 765, 439 to 743, 412 to 700, and 410 to 727; and for the fourth- and fifth-grade subgroup, scores range from 434 to 787, 478 to 777, and 484 to 739.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for indicators of random assignment strata. The values in column 1 (labeled "Enhanced Program") are the observed mean for the members randomly assigned to the enhanced program group. The regular program group values in column 2 are the regression-adjusted means using the observed distribution of the enhanced program group across random assignment strata as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each impact estimate. Statistical significance is indicated by (*) when the pvalue is less than or equal to 5 percent.

The estimated impact effect size of the SAT 10 reading total scaled score is calculated as a proportion of the standard deviation of the regular program group, which is 35.71 based on the analysis sample and the model controlling for demographic characteristics. The standard deviation of a SAT 10 national norming sample with the same grade composition as the study sample is 39.05. For each subtest, the estimated impact effect size is calculated as a proportion of the standard deviation of the regular program group.

There are 7 enhanced program group students and 7 regular program group students who performed at the advanced level on the baseline SAT 10; they are excluded from the prior-achievement subgroup analysis.

[a]The sample consists of second- through fourth-graders only because the spring administration of the test to fifth-graders does not include word study skills.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table F.9

Impact of the Enhanced Reading Program on Student Achievement When Twelve Random Assignment Blocks Are Excluded from the Analysis Sample

Student Achievement Outcome	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
Restricted analysis sample					
SAT 10 reading total scaled scores	588.03	589.27	-1.24	-0.03	0.21
Vocabulary	581.86	582.43	-0.57	-0.01	0.69
Reading comprehension	589.36	590.45	-1.09	-0.03	0.40
Word study skills (grades 2-4)[a]	585.17	587.89	-2.73	-0.07	0.13
Sample size (total = 1,588)	909	679			
Grade subgroups					
Grades 2 and 3					
SAT 10 reading total scaled scores	569.11	571.34	-2.23	-0.06	0.13
Vocabulary	557.30	558.88	-1.58	-0.03	0.47
Reading comprehension	571.13	572.72	-1.58	-0.04	0.42
Word study skills	578.33	583.18	-4.85 *	-0.13	0.02
Sample size (total = 756)	436	320			
Grades 4 and 5					
SAT 10 reading total scaled scores	605.47	605.79	-0.32	-0.01	0.81
Vocabulary	604.49	604.09	0.40	0.01	0.83
Reading comprehension	606.16	606.72	-0.55	-0.01	0.74
Sample size (total = 832)	473	359			
Prior-achievement subgroups					
Students scoring at below basic level					
SAT 10 reading total scaled scores	578.28	576.37	1.91	0.05	0.21
Vocabulary	569.69	567.62	2.07	0.05	0.36
Reading comprehension	580.84	577.99	2.85	0.07	0.14
Word study skills[a]	571.84	572.09	-0.25	-0.01	0.93
Sample size (total = 646)	376	270			
Students scoring at basic level					
SAT 10 reading total scaled scores	592.05	594.27	-2.22	-0.06	0.14
Vocabulary	587.00	588.90	-1.90	-0.04	0.38
Reading comprehension	592.31	594.41	-2.10	-0.05	0.27
Word study skills[a]	589.39	593.81	-4.42	-0.12	0.09
Sample size (total = 763)	435	328			

(continued)

Appendix Table F.9 (continued)

Student Achievement Outcome	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
Students scoring at proficient level					
SAT 10 reading total scaled scores	606.23	612.57	-6.34	-0.18	0.10
Vocabulary	603.81	608.27	-4.46	-0.10	0.44
Reading comprehension	607.34	615.85	-8.51	-0.22	0.11
Word study skills[a]	608.96	607.33	1.63	0.04	0.80
Sample size (total = 166)	91	75			

SOURCE: MDRC calculations are from follow-up results on the Stanford Achievement Test Series, 10th ed. (SAT 10) abbreviated battery.

NOTES: The restricted analysis sample excludes 12 random assignment blocks (grades within centers) because, for each one, the baseline total reading test score difference between the enhanced program and regular program groups is bigger than the overall difference between these groups by more than 1.75 standard deviations.

Based on the SAT 10 national norming sample, total, reading comprehension, vocabulary, and word study skills scaled scores, respectively, have the following possible ranges: for the restricted analysis sample, scores range from 374 to 787, 439 to 777, 412 to 739, and 410 to 740; for the second- and third-grade subgroup, scores range from 374 to 765, 439 to 743, 412 to 700, and 410 to 727; and for the fourth- and fifth-grade subgroup, scores range from 434 to 787, 478 to 777, and 484 to 739.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for indicators of random assignment, baseline reading total scaled score, race/ethnicity, gender, free-lunch status, age, overage for grade, single-adult household, and mother's education. The values in column 1 (labeled "Enhanced Program") are the observed mean for the members randomly assigned to the enhanced program group. The regular program group values in column 2 are the regression-adjusted means using the observed mean covariate values for the enhanced program group as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each impact estimate. Statistical significance is indicated by (*) when the p-value is less than or equal to 5 percent.

The estimated impact effect size of the SAT 10 reading total scaled score is calculated as a proportion of the standard deviation of the regular program group, which is 35.71 based on the full analysis sample. The standard deviation of a SAT 10 national norming sample with the same grade composition as the study sample is 39.05. For each subtest, the estimated impact effect size is calculated as a proportion of the standard deviation of the regular program group.

There are 7 enhanced program group students and 6 regular program group students who performed at the advanced level on the baseline SAT 10; they are excluded from the prior-achievement subgroup analysis.

[a]The sample consists of second- through fourth-graders only because the spring administration of the test to fifth-graders does not include word study skills.

Appendix G

Exploratory Analysis

This appendix lays out the strategy used to investigate possible associations between impacts and characteristics of both the schools housing the after-school program and the implementation of the enhanced after-school program. To explore the interface between the enhanced after-school program strategy and these features, an exploratory correlational analysis was conducted. Because students were not randomly assigned to programs with different school characteristics, this analysis is correlational rather than experimental. As such, the results should not be viewed definitively as causal; the associations that are found could be causal or could purely (or partly) reflect selection bias. Thus, these analyses should be viewed as hypothesis generating, not summative.

In this appendix, the correlational methodology is presented, as is a detailed description of the school characteristic measures used in the analysis.

Analytic Approach

Apart from understanding how impacts may vary with various student characteristics, decision makers may also want to know whether this intervention worked better in particular types of schools or in after-school programs implemented in a particular way. Thus, for the sample of math and reading programs, this part of the analysis explores whether school context characteristics or factors of program implementation were associated with impacts.

Data were collected for the following school characteristics, and their correlation with impacts were examined: the instructional approach of the school-day curricula (available for the math sample but not for the reading sample), how much time is spent in the regular school day on instruction in math or reading, whether the school meet its Adequate Yearly Progress (AYP) goals, what proportion of students in the school receive free or reduced-price lunch, and what is the in-school student-to-teacher ratio. For example, students who are struggling during the school day may benefit from an alternative instructional approach after school. Or additional time in math or reading may have a greater benefit for students who have less time on those topics during the school day. To examine these characteristics, centers were categorized by their regular-school-day curricula (which produced three groups — one with curricula similar to that used after school and two others)[1] as well as categorized by the time spent in the regular school

[1]Note that, for the reading sample, this information is not available.

day on instruction in math or reading (more than 60 minutes per day or less for the math sample, more than 90 minutes per day or less for the reading sample).[2]

Additionally, two factors of program implementation were examined: (1) Did one or more of the instructors teaching the enhanced after-school program leave during the school year? (2) How many days was the enhanced after-school program offered?

The analysis — similar to the approach taken in Bloom, Hill, and Riccio (2001) — examines how the variation of both math and reading impacts is associated with school characteristics across centers.

In particular, this analysis used a two-level hierarchical linear model to estimate how the size of the impact is related to school context inputs. The unit of analysis for Level 1 is the individual student. The unit in Level 2 is the study center. Equations (1) and (2) describe this analytical approach. In this random coefficient model, the size of the center-level impact, β_m, is allowed to vary with the school and the after-school setting experienced by the students.

Level 1

$$Y_{im} = \gamma_1 Y_{-1im} + \sum_{100} \alpha_m Block_{im} + \beta_m T_{im} + \sum_j \delta_j X_{imj} + \varepsilon_{im} \qquad (1)$$

where:

Y_{-1im} = the pretest score for student i in block m before random assignment.[3]

$\sum_j X_{imj}$ = student-level characteristic j for student i from center/block m.

$Block_{im}$ = dummy variable equal to 1 if student i was a member of center/block m, otherwise it is zero.

[2]School administrators were asked how many minutes teachers spend a day teaching math or reading to their students. The responses were not a precise number of minutes, so a continuous measure of minutes is not used. Instead, groups were created around the most common response. For math, 24 percent of schools offer 50 to 60 minutes; 32 percent offer 60 minutes; 28 percent offer 60 to 90 minutes; and the remaining 16 percent offer 90 minutes or more. Thus, for math, the natural split for this subgroup is those offering 60 minutes or less of school-day math instruction and those offering more than 60 minutes. For reading, 20 percent offer, on average, less than 90 minutes (in some schools the amount of time varies by grade); about half (52 percent) offer 90 minutes; and the remaining 28 percent offer more than 90 minutes. Thus, for reading, the natural split is those offering 90 minutes or less and those offering more than 90 minutes.

[3]Pretest scores are scaled scores from the SAT10 tests in reading and math administered in the fall of 2005, before the start of the after-school program. Total scores for each subject are used in the analysis of respective samples.

T_{im} = dummy variable equal to 1 if student *i* was assigned to be part of the experimental group in center/block *m,* otherwise it is zero.

ε_{im} = a student-level random error, assumed to be independently and identically distributed.

Level 2

$$\beta_m = \tau_0 + \tau_1 Group1_m + \tau_2 Group2_m + \tau_3 PERIODlong_m + \tau_4 AYP_m + \tau_5 \%FRL_m + \tau_6 S/T_m + \tau_7 TLEFT_m + \tau_8 TOTDYS_m + \mu_m \quad (2)$$

where:

$Group1_m$ = a dummy equal to 1 if, for the centers implementing Mathletics, the school-day curricula are unit based, which are longer than chapters, and are investigation driven with comparatively fewer practice problems and involving interconnected subproblems, and 0 otherwise.

$Group2_m$ = a dummy equal to 1 if, for the centers implementing Mathletics, the school-day curriculum employs a direct instruction approach organized by lessons with spiraled curriculum, and 0 otherwise.[4]

$PERIODlong_m$ = a dummy equal to 1 if the school-day period in the relevant subject is more than 60 minutes for math or 90 minutes for reading, and 0 otherwise.

AYP_m = a dummy equal to 1 if the school met its AYP requirements in 2005-2006, and 0 otherwise.

$\%FRL_m$ = the percentage of students in school *m* who receive free or reduced-priced lunch centered on the grand mean of all schools in the sample.

S/T_m = a dummy equal to 1 if the student-to-teacher ratio in school *m* is greater than the planned student-to-teacher ratio in the after-school program (13:1 for math).

$TLEFT_m$ = a dummy equal to 1 if one of the instructors teaching the enhanced after-school program left the program during the school year, and 0 otherwise.

[4]In three centers, second-graders used different type of curriculum than the one used in other grades. For these centers, the Group1 and Group2 variables are allowed to vary within school by grade. For example, second-graders within the school may identify with Group2 while the other grades identify with Group1.

$TOTDYS_m$ = the number of days that the enhanced after-school program was offered, centered on the grand mean of all centers in the sample.

μ_m = a center/block-level random error, assumed to be independently and identically distributed.

τ_k (where k = 1, 2, …, 10) is the association between the intervention's impact and school characteristic variable *k,* controlling for other characteristics included in Equation (2). For example, τ_1 is the association of the intervention's impact with having a school-day math curriculum that is unit based, controlling for other characteristics included in Equation (2); and τ_3 is the association of the intervention's impact with having longer periods in school on math or reading, controlling for the other characteristics. If τ_3 is statistically significant and positive, it means that having longer periods in school on math or reading is associated with a bigger program impact.

Appendix H

Service Contrast Subgroups

This appendix shows findings for the *difference* between the after-school academic services received by the enhanced program group and those received by the regular, "business as usual" program group, for subgroups based on student grade level and baseline achievement. The tables present differences in attendance in the after-school program, in hours of instruction received, and in special academic support received from other sources — during the regular school day and outside school.

Appendix Tables H.1 and H.2 present differences for the math program grade-level subgroups and the prior-achievement subgroups, respectively. The difference in hours of academic instruction in math for the second- and third-grade subgroup is 49 hours; for the fourth- and fifth-grade subgroup, it is 48 hours.[1] The difference for the "below basic" and "basic" achievement-level subgroups is 49 hours; for the "proficient" subgroup, it is 46 hours. All these differences are statistically significant.

Findings for the reading program subgroups are presented in Appendix Tables H.3 and H.4. The difference in hours of academic instruction in reading is 51 hours for the second- and third-grade subgroup, and it is 46 hours for fourth- and fifth-graders. The difference for students in the "below basic" achievement level is 43 hours; at the "basic" level, it is 51 hours; and at the "proficient" level, it is 53 hours. All these differences are statistically significant.

Overall, for both measures of attendance in the after-school program, in all but one case, the findings for reading and math subgroups based on student grade level and baseline achievement are similar to those found for the analysis sample, with the same pattern of somewhat greater attendance among the enhanced program group.[2]

[1] In addition, tests found that there are no significantly different patterns of service contrast by grade level within the younger and older subgroups.

[2] One subgroup — the reading students scoring at the "proficient level" — has a negative impact estimate of –1.1 (p-value = 0.81) for number of days attended.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table H.1

Attendance of Students in the Math Analysis Sample, by Grade Subgroup

Attendance Measure	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
Grades 2 and 3					
Attendance in after-school program[a]					
Number of days attended	74.65	62.47	12.18 *	0.37	0.00
Total hours of math instruction received[b]	58.07	9.03	49.04 *	2.78	0.00
Math support from other sources					
Out-of-school math class or tutoring[c]					
Students receiving instruction (%)	35.65	24.10	11.54 *	0.29	0.00
Number of days per week[d]	1.21	0.69	0.52 *	0.37	0.00
Regular school day[e]					
Students receiving special support (%)	2.21	2.19	0.02	0.05	0.40
Minutes per week of individualized help	42.60	43.56	-0.96	-0.01	0.79
Sample size (total = 971)	533	438			
Grades 4 and 5					
Attendance in after-school program[a]					
Number of days attended	72.29	60.11	12.18 *	0.37	0.00
Total hours of math instruction received[b]	56.30	8.22	48.08 *	2.73	0.00
Math support from other sources					
Out-of-school math class or tutoring[c]					
Students receiving instruction (%)	21.90	17.75	4.14	0.10	0.07
Number of days per week[d]	0.73	0.50	0.23 *	0.17	0.01
Regular school day[e]					
Students receiving special support (%)	2.27	2.30	-0.04	-0.08	0.15
Minutes per week of individualized help	56.94	55.13	1.81	0.03	0.89
Sample size (total = 990)	548	442			

(continued)

SOURCES: MDRC calculations are from the Evaluation of Academic Instruction in After-School Programs attendance records, student survey responses, and regular-school-day teacher survey responses.

NOTES: The estimated impacts are regression-adjusted using ordinary least squares, controlling for indicators of random assignment, baseline math total scaled score, race/ethnicity, gender, free-lunch status, age, overage for grade, single-adult household, and mother's education. The values in column 1 (labeled "Enhanced Program") are the observed mean for the members randomly assigned to the enhanced program group. The regular program group values in column 2 are the regression-adjusted means using the observed mean covariate values for the enhanced program group as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

Appendix Table H.1 (continued)

A two-tailed t-test was applied to each impact estimate. Statistical significance is indicated by (*) when the p-value is less than or equal to 5 percent.

The estimated impact effect size for each measure is calculated as a proportion of the standard deviation of the regular program group.

[a]Attendance in the after-school program is based on the days the enhanced program operated.

[b]Students in the enhanced classes received 45 minutes of instruction (and 60 minutes in one site that met only three days a week) on the days they were present. Total hours is calculated for these students by multiplying each student's total days of attendance by 45 (or 60 in the one site).

Students in the regular program group were not supposed to receive any structured instruction. However, some regular program staff indicated on the survey that they provide structured academic instruction. Total hours is calculated for these students by multiplying the total number of days attended by 45, then by the proportion of regular program staff within the center who reported providing structured instruction. If no regular program staff in a center indicated that they provide structured instruction, then total hours for these students in that center is zero. If no regular program staff in a center answered this question, this calculation could not be performed for these students. Calculated as such, the sample sizes for the regular program group are 379 for the second- and third-grade subgroup and 391 for the fourth- and fifth-grade subgroup.

[c]This information comes from student survey responses to questions for each day of the week that ask, "Do you go somewhere else for a math class or to be tutored in math?" These calculations are based on a smaller sample than the reported analysis sample by the number of students who did not complete a survey. For the second- and third-grade subgroup, the sample size is 533 for the enhanced program group and 437 for the regular program group. For the fourth- and fifth-grade subgroup, the sample size is 548 for the enhanced program group and 442 for the regular program group.

[d]Students who responded that they do not receive math support from other out-of-school sources are included in these averages.

[e]This information comes from regular-school-day teacher survey responses. "Special support" refers to special support in math during the school day (that is, pull-out tutoring, remedial math assistance, assigned to a computer-assisted lab, and so on). "Individualized help" refers to individual help from the teacher or an aide with a task or answering a question. Teachers who responded that they did not provide support may or may not have responded that they provided minutes of individualized help. Thus, average minutes includes responses for all students, not just those who received special support.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table H.2

Attendance of Students in the Math Analysis Sample, by Prior-Achievement Subgroup

Attendance Measure	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
Students scoring at below basic level					
Attendance in after-school program[a]					
Number of days attended	66.80	55.66	11.14 *	0.34	0.00
Total hours of math instruction received[b]	52.63	4.03	48.60 *	2.76	0.00
Math support from other sources					
Out-of-school math class or tutoring[c]					
Students receiving instruction (%)	41.42	34.47	6.95	0.17	0.10
Number of days per week[d]	1.50	1.08	0.42 *	0.30	0.01
Regular school day[e]					
Students receiving special support (%)	2.37	2.43	-0.06	-0.13	0.19
Minutes per week of individualized help	58.03	68.20	-10.17	-0.15	0.08
Sample size (total = 467)	239	228			
Students scoring at basic level					
Attendance in after-school program[a]					
Number of days attended	74.50	60.04	14.47 *	0.44	0.00
Total hours of math instruction received[b]	57.91	8.55	49.36 *	2.80	0.00
Math support from other sources					
Out-of-school math class or tutoring[c]					
Students receiving instruction (%)	27.45	20.46	6.99 *	0.17	0.00
Number of days per week[d]	0.84	0.57	0.27 *	0.20	0.00
Regular school day[e]					
Students receiving special support (%)	2.24	2.23	0.01	0.02	0.72
Minutes per week of individualized help	53.52	52.01	1.51	0.02	0.90
Sample size (total = 1,055)	612	443			
Students scoring at proficient level					
Attendance in after-school program[a]					
Number of days attended	77.83	70.79	7.04 *	0.22	0.01
Total hours of math instruction received[b]	60.25	14.64	45.61 *	2.59	0.00
Math support from other sources					
Out-of-school math class or tutoring[c]					
Students receiving instruction (%)	18.81	12.09	6.72	0.17	0.06
Number of days per week[d]	0.75	0.37	0.38 *	0.27	0.01

(continued)

Appendix Table H.2 (continued)

Attendance Measure	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
Regular school day[e]					
Students receiving special support (%)	2.10	2.08	0.02	0.05	0.55
Minutes per week of individualized help	33.44	30.75	2.69	0.04	0.63
Sample size (total = 380)	202	178			

SOURCES: MDRC calculations are from the Evaluation of Academic Instruction in After-School Programs attendance records, student survey responses, and regular-school-day teacher survey responses.

NOTES: The estimated impacts are regression-adjusted using ordinary least squares, controlling for indicators of random assignment, baseline math total scaled score, race/ethnicity, gender, free-lunch status, age, overage for grade, single-adult household, and mother's education. The values in column 1 (labeled "Enhanced Program") are the observed mean for the members randomly assigned to the enhanced program group. The regular program group values in column 2 are the regression-adjusted means using the observed mean covariate values for the enhanced program group as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each impact estimate. Statistical significance is indicated by (*) when the p-value is less than or equal to 5 percent.

The estimated impact effect size for each measure is calculated as a proportion of the standard deviation of the regular program group.

[a]Attendance in the after-school program is based on the days the enhanced program operated.

[b]Students in the enhanced classes received 45 minutes of instruction (and 60 minutes in one site that met only three days a week) on the days they were present. Total hours is calculated for these students by multiplying each student's total days of attendance by 45 (or 60 in the one site).

Students in the regular program group were not supposed to receive any structured instruction. However, some regular program staff indicated on the survey that they provide structured academic instruction. Total hours is calculated for these students by multiplying the total number of days attended by 45, then by the proportion of regular program staff within the center who reported providing structured instruction. If no regular program staff in a center indicated that they provide structured instruction, then total hours for these students in that center is zero. If no regular program staff in a center answered this question, this calculation could not be performed for these students. Calculated as such, the sample sizes for the regular program group are 181 for the group of students scoring at the below basic level, 397 for the group of students scoring at the basic level, and 164 for the group of students scoring at the proficient level.

[c]This information comes from student survey responses to questions for each day of the week that ask, "Do you go somewhere else for a math class or to be tutored in math?" These calculations are based on a smaller sample than the reported analysis sample by the number of students who did not complete a survey. For the group of students scoring at the below basic level, the sample size is 239 for the enhanced program group and 227 for the regular program group. For the group of students scoring at the basic level, the sample size is 612 for the enhanced program group and 443 for the regular program group. For the group of students scoring at the proficient level, the sample size is 202 for the enhanced program group and 178 for the regular program group.

[d]Students who responded that they do not receive math support from other out-of-school sources are included in these averages.

[e]This information comes from regular-school-day teacher survey responses. "Special support" refers to special support in math during the school day (that is, pull-out tutoring, remedial math assistance, assigned to a computer-assisted lab, and so on). "Individualized help" refers to individual help from the teacher or an aide with a task or answering a question. Teachers who responded that they did not provide support may or may not have responded that they provided minutes of individualized help. Thus, average minutes includes responses for all students, not just those who received special support.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table H.3

Attendance of Students in the Reading Analysis Sample, by Grade Subgroup

Attendance Measure	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
Grades 2 and 3					
Attendance in after-school program[a]					
Number of days attended	73.34	64.61	8.73 *	0.25	0.00
Total hours of reading instruction received[b]	57.14	6.04	51.10 *	2.89	0.00
Reading support from other sources					
Out-of-school reading class or tutoring[c]					
Students receiving instruction (%)	45.68	36.39	9.29 *	0.20	0.00
Number of days per week[d]	1.37	0.95	0.42 *	0.29	0.00
Regular school day[e]					
Students receiving special support (%)	2.46	2.41	0.05	0.09	0.10
Minutes per week of individualized help	73.03	69.55	3.47	0.02	0.58
Sample size (total = 912)	524	388			
Grades 4 and 5					
Attendance in after-school program[a]					
Number of days attended	67.33	62.73	4.60 *	0.13	0.01
Total hours of reading instruction received[b]	52.86	7.21	45.65 *	2.58	0.00
Reading support from other sources					
Out-of-school reading class or tutoring[c]					
Students receiving instruction (%)	31.59	26.24	5.35	0.12	0.06
Number of days per week[d]	0.88	0.61	0.28 *	0.19	0.00
Regular school day[e]					
Students receiving special support (%)	2.36	2.37	-0.01	-0.02	0.66
Minutes per week of individualized help	100.37	98.73	1.64	0.01	0.83
Sample size (total = 916)	524	392			

(continued)

SOURCES: MDRC calculations are from the Evaluation of Academic Instruction in After-School Programs attendance records, student survey responses, and regular-school-day teacher survey responses.

NOTES: The estimated impacts are regression-adjusted using ordinary least squares, controlling for indicators of random assignment, baseline reading total scaled score, race/ethnicity, gender, free-lunch status, age, overage for grade, single-adult household, and mother's education. The values in column 1 (labeled "Enhanced Program") are the observed mean for the members randomly assigned to the enhanced program group. The regular program group values in column 2 are the regression-adjusted means using the observed mean covariate values for the enhanced program group as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

Appendix Table H.3 (continued)

A two-tailed t-test was applied to each impact estimate. Statistical significance is indicated by (*) when the p-value is less than or equal to 5 percent.

The estimated impact effect size for each measure is calculated as a proportion of the standard deviation of the regular program group.

[a]Attendance in the after-school program is based on the days the enhanced program operated.

[b]Students in the enhanced classes received 45 minutes of instruction (and 60 minutes in one site that met only three days a week) on the days they were present. Total hours is calculated for these students by multiplying each student's total days of attendance by 45 (or 60 in the one site).

Students in the regular program group were not supposed to receive any structured instruction. However, some regular program staff indicated on the survey that they provide structured academic instruction. Total hours is calculated for these students by multiplying the total number of days attended by 45, then by the proportion of regular program staff within the center who reported providing structured instruction. If no regular program staff in a center indicated that they provide structured instruction, then total hours for these students in that center is zero. If no regular program staff in a center answered this question, this calculation could not be performed for these students. Calculated as such, the sample size for the regular program group is 299 for the second- and third-grade subgroup and 304 for the fourth- and fifth-grade subgroup.

[c]This information comes from student survey responses to questions for each day of the week that ask, "Do you go somewhere else for a reading class or to be tutored in reading?" These calculations are based on a smaller sample than the reported analysis sample by the number of students who did not complete a survey. For the second- and third-grade subgroup, the sample size is 521 for the enhanced program group and 386 for the regular program group. For the fourth- and fifth-grade subgroup, the sample size is 516 for the enhanced program group and 386 for the regular program group.

[d]Students who responded that they do not receive reading support from other out-of-school sources are included in these averages.

[e]This information comes from regular-school-day teacher survey responses. "Special support" refers to special support in reading during the school day (that is, pull-out tutoring, Reading Recovery, assigned to a computer-assisted lab, and so on). "Individualized help" refers to individual help from the teacher or an aide with a task or answering a question. Teachers who responded that they did not provide support may or may not have responded that they provided minutes of individualized help. Thus, average minutes includes responses for all students, not just those who received special support.

The Evaluation of Academic Instruction in After-School Programs

Appendix Table H.4

ttendance of Students in the Reading Analysis Sample, by Prior-Achievement Subgrou

Attendance Measure	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
Students scoring at below basic level					
Attendance in after-school program[a]					
Number of days attended	63.39	60.14	3.25	0.09	0.12
Total hours of reading instruction received[b]	50.03	7.25	42.78 *	2.42	0.00
Reading support from other sources					
Out-of-school reading class or tutoring[c]					
Students receiving instruction (%)	40.05	38.85	1.20	0.03	0.74
Number of days per week[d]	1.16	1.02	0.13	0.09	0.27
Regular school day[e]					
Students receiving special support (%)	2.51	2.53	-0.02	-0.04	0.62
Minutes per week of individualized help	111.29	114.49	-3.20	-0.02	0.75
Sample size (total = 736)	437	299			
Students scoring at basic level					
Attendance in after-school program[a]					
Number of days attended	75.30	65.60	9.71 *	0.28	0.00
Total hours of reading instruction received[b]	58.63	7.34	51.29 *	2.90	0.00
Reading support from other sources					
Out-of-school reading class or tutoring[c]					
Students receiving instruction (%)	38.80	27.42	11.38 *	0.25	0.00
Number of days per week[d]	1.14	0.68	0.46 *	0.32	0.00
Regular school day[e]					
Students receiving special support (%)	2.36	2.33	0.03	0.06	0.30
Minutes per week of individualized help	72.52	69.19	3.33	0.02	0.59
Sample size (total = 877)	501	376			
Students scoring at proficient level					
Attendance in after-school program[a]					
Number of days attended	75.09	76.19	-1.10	-0.03	0.81
Total hours of reading instruction received[b]	58.15	5.36	52.79 *	2.98	0.00
Reading support from other sources					
Out-of-school reading class or tutoring[c]					
Students receiving instruction (%)	32.04	18.91	13.13	0.28	0.12
Number of days per week[d]	0.94	0.46	0.48	0.33	0.07

(continued)

Appendix Table H.4 (continued)

Attendance Measure	Enhanced Program	Regular Program	Estimated Impact	Estimated Impact Effect Size	P-Value for the Estimated Impact
Regular school day[e]					
Students receiving special support (%)	2.23	2.20	0.03	0.07	0.62
Minutes per week of individualized help	54.24	52.14	2.10	0.01	0.85
Sample size (total = 201)	103	98			

SOURCES: MDRC calculations are from the Evaluation of Academic Instruction in After-School Programs attendance records, student survey responses, and regular-school-day teacher survey responses.

NOTES: The estimated impacts are regression-adjusted using ordinary least squares, controlling for indicators of random assignment, baseline reading total scaled score, race/ethnicity, gender, free-lunch status, age, overage for grade, single-adult household, and mother's education. The values in column 1 (labeled "Enhanced Program") are the observed mean for the members randomly assigned to the enhanced program group. The regular program group values in column 2 are the regression-adjusted means using the observed mean covariate values for the enhanced program group as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each impact estimate. Statistical significance is indicated by (*) when the p-value is less than or equal to 5 percent.

The estimated impact effect size for each measure is calculated as a proportion of the standard deviation of the regular program group.

[a]Attendance in the after-school program is based on the days the enhanced program operated.

[b]Students in the enhanced classes received 45 minutes of instruction (and 60 minutes in one site that met only three days a week) on the days they were present. Total hours is calculated for these students by multiplying each student's total days of attendance by 45 (or 60 in the one site).

Students in the regular program group were not supposed to receive any structured instruction. However, some regular program staff indicated on the survey that they provide structured academic instruction. Total hours is calculated for these students by multiplying the total number of days attended by 45, then by the proportion of regular program staff within the center who reported providing structured instruction. If no regular program staff in a center indicated that they provide structured instruction, then total hours for these students in that center is zero. If no regular program staff in a center answered this question, this calculation could not be performed for these students. Calculated as such, the sample size for the regular program group is 242 for the group of students scoring at the below basic level, 285 for the group of students scoring at the basic level, and 73 for the group of students scoring at the proficient level.

[c]This information comes from student survey responses to questions for each day of the week that ask, "Do you go somewhere else for a reading class or to be tutored in reading?" These calculations are based on a smaller sample than the reported analysis sample by the number of students who did not complete a survey. For the group of students scoring at the below basic level, the sample size is 427 for the enhanced program group and 297 for the regular program group. For the group of students scoring at the basic level, the sample size is 500 for the enhanced program group and 371 for the regular program group. For the group of students scoring at the proficient level, the sample size is 103 for the enhanced program group and 97 for the regular program group.

[d]Students who responded that they do not receive reading support from other out-of-school sources are included in these averages.

[e]This information comes from regular-school-day teacher survey responses. "Special support" refers to special support in reading during the school day (that is, pull-out tutoring, Reading Recovery, assigned to a computer-assisted lab, and so on). "Individualized help" refers to individual help from the teacher or an aide with a task or answering a question. Teachers who responded that they did not provide support may or may not have responded that they provided minutes of individualized help. Thus, average minutes includes responses for all students, not just those who received special support.

References

Arbreton, A. J., Goldsmith, J., and Sheldon, J. (2005). *Launching Literacy in After-School Programs: Early Lessons from the CORAL Initiative.* Philadelphia: Public/Private Ventures.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A New and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B*(57): 1289-1300.

Bloom, H. S., Hill, C. J., and Riccio, J. (2001). *Modeling the Performance of Welfare-to-Work Programs: The Effects of Program Management and Services, Economic Environment, and Client Characteristics.* MDRC Working Papers on Research Methodology. New York: MDRC. Web site: www.mdrc.org.

Bodilly, S., and Beckett, M. K. (2005). *Making Out-of-School Time Matter: Evidence for an Action Agenda.* Santa Monica, CA: RAND Corporation.

Britsch, B., Martin, N., Stuczynski, A., Tomala, B., and Tucci, P. (2005). *Literacy in Afterschool Programs: Literature Review.* Portland, OR: Northwest Regional Educational Laboratory.

DeVellis, R. F. (2003). *Scale Development: Theory and Applications.* In Applied Social Research Methods Series (vol. 26). Thousand Oaks, CA: Sage Publications.

Dynamic Indicators of Basic Early Literacy Skills (DIBELS). (2007a). *DIBELS Data System.* Retrieved September 7, 2007, from http://dibels.uoregon.edu/index.php.

Dynamic Indicators of Basic Early Literacy Skills (DIBELS). (2007b). *Oral Reading Fluency, Retell Fluency.*
Retrieved September 7 , 2007, from http://dibels.uoregon.edu/measures/orf.php.

Dynamic Indicators of Basic Early Literacy Skills (DIBELS). (2007c). *Nonsense Word Fluency.* Retrieved September 7, 2007, from http://dibels.uoregon.edu/measures/nwf.php.

Dynarski, M., James-Burdumy, S., Moore, M., Rosenberg, L., Deke, J., and Mansfield, W. (2004). *When Schools Stay Open Late: The National Evaluation of the 21st Century Community Learning Centers Program: New Findings*. U.S. Department of Education, National Center for Education Evaluation and Regional Assistance. Washington, DC: U.S. Government Printing Office.

Dynarski, M., Moore, M., Mullens, J., Gleason, P., James-Burdumy, S., Rosenberg, L., Pistorino, C., Silva, T., Deke, J., Mansfield, W., Heaviside, S., and Levy, D. (2003). *When Schools Stay Open Late: The National Evaluation of the 21st Century Community Learning Centers Program, First-Year Findings.* Report submitted to the U.S. Department of Education. Princeton, NJ: Mathematica Policy Research, Inc.

Ferrandino, Vincent L. (2007). *A New Day for Learning.* Denver, CO: Time, Learning, and Afterschool Task Force.

Grossman, J. B., Campbell, M., and Raley, R. (2007). *Quality After-School Time: What Instructors Can Do to Enhance Learning.* Philadelphia: Public/Private Ventures.

Harcourt Assessment. *Stanford Achievement Test Series, Tenth Edition – Abbreviated Battery, Areas of Assessment.* (n.d.) San Antonio, TX: Harcourt Assessment, Inc. Retrieved September 9, 2007, from http://harcourtassessment.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=SAT10A&Mode=summary&Leaf=SAT10A_2.

Harcourt Assessment. *Stanford Achievement Test Series, Tenth Edition – Spring Multilevel Norms Book.* (2003). San Antonio, TX: Harcourt Assessment, Inc.

Harcourt Assessment. *Stanford Achievement Test Series, Tenth Edition – Technical Data Report.* (2004). San Antonio, TX: Harcourt Assessment, Inc.

Kane, T. J. (2004, January 16). *The Impact of After-School Programs: Interpreting the Results of Four Recent Evaluations.* William T. Grant Foundation Working Paper. New York: William T. Grant Foundation.

Miller, B. M. (2003). *Critical Hours: Afterschool Programs and Educational Success*. Brookline, MA: Miller Midzik Research Associates.

Miller, B. (2006). *Pathways to Success: What Counts in After School Programs.* Wellesley, MA: National Institute for Out-of School Time.

Naftzger, N., Kaufman, S., Margolin, J., and Ali, A. (2006, July). *An Overview of the 21st CCLC Program: 2004-05* (pp. 34-40). Naperville, IL: Learning Point Associates.

National Center for Education Statistics, Common Core of Data (CCD). (n.d.) *Public Elementary/Secondary School Universe Survey Data: 2005-06.* Retrieved July 26, 2007, from http://nces.ed.gov/ccd/.

National Center for Education Statistics, Common Core of Data (CCD). (n.d.) *Identification of Rural Locales.* Retrieved August 9, 2007, from http://nces.ed.gov/ccd/Rural_Locales.asp.

National Mathematics Advisory Panel. (2008). *Foundations for Success: The Final Report of the National Mathematics Advisory Panel.* U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

National Reading Panel. (2000). *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Literature on Reading and Its Implications for Reading Instruction.* Bethesda, MD: National Institute of Child Health and Human Development.

National Research Council Institute of Medicine. (2004). *Engaging Schools: Fostering High School Students' Motivation to Learn.* Washington, DC: National Academic Press.

Salvia, J., and Ysseldyke, J. E. (2001). *Assessment* (8th ed.). Boston: Houghton Mifflin Company.

Scantron Corporation. (2005). *Above and Beyond: Applying Adaptive Technology to Diagnose Student Performance and Progress.* Web site: http://www.scantron.com/downloads/Performance_Series_White_Paper.pdf.

Shepard, L. A. The Role of Classroom Assessment in Teaching and Learning. (2001). In Virginia Richardson (Ed.), *The Handbook of Research on Teaching* (8th ed., pp. 1066-1101). Washington DC: American Education Research Association.

Slavin, R. E., and Madden, N. A. (1999). *Disseminating Success for All: Lessons for Policy and Practice.* Center for Research on the Education of Students Placed At Risk (CRESPAR) Report No. 30. Web site: www.csos.jhu.edu/crespar/.

Success for All. (2004). *Adventure Island Reading Intervention for Grades 1 to 5: Targeted Reading Intervention for Struggling Students.* Marketing brochure prepared by the Success for All Foundation to describe the program.

U.S. Department of Education, 21st Century Community Learning Centers. (n.d.) *Program Description*.
Retrieved September 7, 2007, from http://www.ed.gov/programs/21stcclc/index.html.